Bayesian Parameter Estimation

STAT 489-01: Bayesian Methods of Data Analysis *

Spring Semester 2017

Contents

Tuesday 24 January 2017
– Refer to Chapter 1 of Gelman and/or Chapter 4 of Bolstad

0 Preliminaries

0.1 Administrata

- Introductions!
- Outcome of clipboard survey on mathematical and computer background.
- Syllabus
- Instructor's name (Whelan) rhymes with "wailin'".
- Books. Note that we'll be following the nominal required text fairly loosely, and students are encouraged to refer to whatever resources best mesh with their learning style. Note also that there's a new edition of Bolstad out, so you might be able to get a better deal on a used copy the 2nd edition (although that won't have some of the later chapters, notably Chapter 20 on computational methods).
  - Gelman et al, *Bayesian Data Analysis*, 3rd edition. This is the nominal textbook for the class, and we'll be using it to set the approximate sequence of topics, notation, etc., and taking exercises from it. Its level of theoretical mathematical detail is a bit too high for this course at times, however.
  - Bolstad, *Introduction to Bayesian Statistics*, 2nd or 3rd edition. This book probably has the best perspective of the recommended texts, but it is organized by specific problem rather than topics, which is why we're not using it as the text. Also, it's trying to do double duty and stand in as an intro stats text for people who haven't taken a course in classical statistics, so it moves a little slowly.
  - McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 1st edition. This brand-new book provides an accessible hands-on introduction to some Bayesian techniques. It's designed for researchers and graduate students in social and natural sciences, so it assumes a little less math than this course, but it may be a very good learning tool.[1]
  - Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan* this book is very informal and presents a lot of the concepts by example and illustration.
  - Jaynes, *Probability Theory: the Logic of Science*. This is a sort of Bayesian manifesto (written by a Physicist), but it's got a lot of interesting bits in it, as well as a clear illustration of a particular Bayesian philosophy (to which I am sympathetic) and some amusingly snarky quotes. It's also got a lot of mathematical detail on Bayesian concepts (e.g., a demonstration that you can derive probability as an obvious extension of logic).
  - Sivia with Skilling, *Data Analysis: A Bayesian Tutorial*. This is a short book which gives a limited but insightful introduction to some simple Bayesian concepts and methods.
- Course website: `http://ccrg.rit.edu/~whelan/STAT-489/`
  - Will contain links to notes and problem sets; course calendar is probably the most useful.

[1] This assumption is not entirely inconvenient. Jaynes, when describing the audience for his book, says: "A previous acquaintance with probability and statistics is not necessary; indeed, a certain amount of innocence in this area may be desirable, because there will be less to unlearn."

  - Course calendar: *tentative* timetable for course; will evolve as the semester progresses.

- Course work:
  - It's important to use the outside reading to complement the presentation in the lectures. I'll try to keep the notes up-to-date (also useful for me!) but I can't promise they won't be a few days behind, so it's a good idea to get someone's notes if you have to miss class.
  - There will be quasi-weekly homeworks. Collaboration is allowed an encouraged, but please turn in your own work, as obviously identical homeworks may not receive credit.
  - We'll have a longer-term project towards the end of the semester.
  - There will be two prelim exams, in class, and one cumulative final exam.

- Grading:

  25% Homework (Problem Sets and Final Project)
  20% First Prelim Exam
  20% Second Prelim Exam
  35% Final Exam

  You'll get a separate grade on the "quality point" scale (e.g., 3.1667–3.5 is the B+ range) for each of these five components; course grade is weighted average.

0.2 Outline

1. Bayesian Parameter Estimation (Gelman Chapters 1-5)
2. Bayesian Model Comparison (Gelman Chapters 6-9)
3. Advanced Computational Techniques (Gelman Chapters 10-13)

1 Bayesian Probability

The various techniques of Bayesian data analysis are motivated by a few basic principles, so once we spell out what it is we're trying to calculate, most of the rest is just details on how to evaluate those quantities. (In contrast, classical statistical techniques often require a number of arbitrary choices to define the statistical tests of interest.)[2]

There are two major features that set the Bayesian interpretation of probability apart from the usual frequentist interpretation:

1. A probability can in principle be assigned to any proposition which could be true or false. I.e., the allowable set of "events" consists not only of the outcomes of repeatable experiments but of all logical propositions. Notably, that includes statements about the correctness of models and parameters having particular values or ranges of values.
2. **All probabilities are conditional.** I.e., a probability $\Pr(A|I)$ is always defined in the context of some underlying information (or state of knowledge) I. This is sometimes referred to as "subjective" because observers with different states of knowledge I_1 and I_2 will assign different probabilities to the same event: $\Pr(A|I_1) \neq \Pr(A|I_2)$.

1.1 Logic and Probability

The interpretation of probability which makes the most sense in the Bayesian context is that of extended logic.[3] If a proposition A is known to be definitely true in light of information I, then

[2] In fact, the reason why Bolstad is structured as a series of applications is that there really aren't many different underlying formalisms to develop.

[3] Hence the title of Jaynes's book, *Probability Theory: The Logic of Science*.

$\Pr(A|I) = 1$. If it's definitely false, $\Pr(A|I) = 0$. If we're uncertain about its truth or falsehood, $0 < \Pr(A|I) < 1$ quantifies our degree of confidence that A is true given I. Fundamentally, in this framework, **probability is a measure of uncertainty** or lack of knowledge, not necessarily of inherent randomness.

If A represents the outcome of an experiment which we could somehow arrange to repeat under identical circumstances, then $\Pr(A|I)$ will be approximately equal to the long-term frequency of the event A. I.e., if we do some large number N of repetitions of the experiment, at the beginning of which we recreate the situation described by I, the approximate number of experiments in which A will turn out to be true is $N \times \Pr(A|I)$. In the classical or "frequentist" approach to statistics, this is the only sort of event to which we're allowed to assign a probability, but in the more general Bayesian framework we are free to assign probabilities to any logical proposition.

It's a bit remarkable that the conventional formulation of probability in terms of set theory is basically equivalent to a formulation in terms of logic. In particular, the three basic operations used to combine logical propositions can be understood in terms of logic.

- The complement A^C or $\overline{A}$ is logical negation "not A", and can also be written as A' or $\neg A$.
- The intersection $A \cap B$ is a logical conjunction "A and B", which can also be written $A \wedge B$ or A, B. This is the notation implicitly used by Gelman: $\Pr(A, B|I)$ is the probability that both A and B are true, given I.
- The union $A \cup B$ is a logical "disjunction" "A or B", which can also be written $A \vee B$ or (somewhat counterintuitively) $A + B$.

There are basic rules of probability corresponding to these logical operations:

- $\Pr(A|I) + \Pr(\overline{A}|I) = 1$.
- The product rule: $\Pr(A, B|I) = \Pr(A|B, I) \Pr(B|I)$.
- The sum rule: if A and B are mutually exclusive, i.e., if $\Pr(A, B|I) = 0$, then $\Pr(A \vee B|I) = \Pr(A|I) + \Pr(B|I)$.

Note that in this approach, it's the product rule which is fundamental, although in the classical framework, it's a rearrangement of the definition of conditional probability $\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$, which we now understand as

$$\Pr(A|B, I) = \frac{\Pr(A, B|I)}{\Pr(B|I)} \tag{1.1}$$

1.2 Bayes's Theorem

Because the logical "and" operation is symmetrical, i.e., A, B is equivalent to B, A, we can write the product rule in two different ways:

$$\Pr(A, B|I) = \Pr(A|B, I) \Pr(B|I) = \Pr(B|A, I) \Pr(A|I) \tag{1.2}$$

this can be rearranged into *Bayes's Theorem*, which says that

$$\Pr(A|B, I) = \frac{\Pr(B|A, I) \Pr(A|I)}{\Pr(B|I)} \tag{1.3}$$

which is incredibly useful when you naturally know $\Pr(B|A, I)$ but would like to know $\Pr(A|B, I)$. For instance, suppose A refers to "I have terrible-disease-of-the-year (TDY)", B refers to "I test positive for TDY", and I represents the information that I had no extra risk factors or symptoms for TDY but was routinely tested, In fact, rather than try to remember which is A and which is B, let's use "sick" for A, "healthy" for $\overline{A}$, "pos" for B and "neg" for $\overline{B}$. Now suppose 0.1% of people in such

a group have TDY, the test has a 2% false positive rate (2% of people without TDY will test positive for it) and a 1% false negative rate (1% of people with TDY will test negative for it). This information tells us that:

- $\Pr(\text{sick}|I) = 0.001$ so $\Pr(\text{healthy}|I) = 0.999$.
- $\Pr(\text{neg}|\text{sick}, I) = 0.01$ so $\Pr(\text{pos}|\text{sick}, I) = 0.99$.
- $\Pr(\text{pos}|\text{healthy}, I) = 0.02$ so $\Pr(\text{neg}|\text{healthy}, I) = 0.98$.

Additionally, since $\text{pos} = (\text{pos}, \text{sick}) \vee (\text{pos}, \text{healthy})$,

$$\begin{aligned}\Pr(\text{pos}|I) &= \Pr(\text{pos}, \text{sick}|I) + \Pr(\text{pos}, \text{healthy}|I) \\ &= \Pr(\text{pos}|\text{sick}, I)\Pr(\text{sick}|I) \\ &\quad + \Pr(\text{pos}|\text{healthy}, I)\Pr(\text{healthy}|I) \\ &= 0.99 \times 0.001 + 0.02 \times 0.999 = 0.00099 + 0.01998 \\ &= 0.02097\end{aligned} \tag{1.4}$$

We can then use Bayes's theorem to show that

$$\Pr(\text{sick}|\text{pos}, I) = \frac{0.00099}{0.02097} \approx 0.04721 \tag{1.5}$$

I.e., if I test positive for TDY, I have about a 4.7% chance of actually having the disease. This is a lot less than $\Pr(\text{pos}|\text{sick}, I)$, which is 99%!

1.2.1 The Hypothetical Population

There's a nice illustration, among other places, at
`http://yudkowsky.net/rational/bayes`
that it's sometimes easier to follow what's happening in this argument by considering a population of individuals described by the probabilities above. So if there are a million people, and one in a thousand have TDY, that's 1,000. The other 999,000 do not have it. The 2% false positive rate means that of the 999,000 healthy individuals, 2% of them, or 19,980, will test positive. The other 979,020 will test negative. The 1% false negative rate means that of the 1,000 sick individuals, ten will test negative and the other 990 will test positive. So let's collect this into a table:

	Positive	Negative	Total
Sick	990	10	1,000
Healthy	19,980	979,020	999,000
Total	20,970	979,030	1,000,000

(Of course, if we choose a *sample* of a million individuals out of a larger population, we won't expect to get exactly this number of results, but the representative population is still useful conceptual construct.)

Translating from numbers in this hypothetical population, we can confirm that it captures the input information:

$$P(\text{sick}) = \frac{1,000}{1,000,000} = .001 \tag{1.6a}$$

$$P(\text{positive}|\text{healthy}) = \frac{19,980}{999,000} = .02 \tag{1.6b}$$

$$P(\text{negative}|\text{sick}) = \frac{10}{1,000} = .01 \tag{1.6c}$$

But now we can also calculate what we want, the conditional probability of being sick given a positive result. That is the fraction of the total number of individuals with positive test results that are in the "sick *and* positive" category:

$$P(\text{sick}|\text{positive}) = \frac{990}{20,970} \approx .04721 \tag{1.7}$$

or about 4.7%, as before.

A few comments about the hypothetical population construct. In the disease testing example there really is a population of individuals that we can talk about, but the whole point of Bayesian probability is that we can do all the same calculations even when the proposition in question doesn't represent a draw from a population. For example, before the New Horizons spacecraft visited Pluto, it was unknown whether it was larger or smaller (in radius) than Eris, the most massive Trans-Neptunian Object. A reasonable summary of the available information would have been to assign a probability to the proposition "Pluto is larger than Eris". But there is not an ensemble of Pluto-Eris pairs from which ours is drawn. There is one Pluto and one Eris, and we now know that Pluto is actually larger. On the other hand, even when a probability does not represent a draw from a population, it can be a useful technique to simulate draws from an imaginary population described by the same distribution.[4] In fact this is the basis for Markov Chain Monte Carlo and other Monte Carlo methods which interact with a probability distribution by sampling from it.

1.2.2 Bayes's Theorem and Data Analysis

In the context of observational science, Bayes's theorem is most commonly applied to a situation where H is a hypothesis which I'd like to evaluate and D is a particular set of data I've collected. It's usually straightforward to work out $\Pr(D|H, I)$, the probability of observing a particular set of data values given a model, but I generally want to answer the question, what is my degree of belief in the hypothesis H after the observation. The answer, according to Bayes's Theorem, is

$$\Pr(H|D, I) = \frac{\Pr(D|H, I)\Pr(H|I)}{\Pr(D|I)} \tag{1.8}$$

The various quantities in this expression have standard names:

- $\Pr(H|I)$ is the prior probability of the hypothesis H, i.e., the probability we'd assign based on the background information I, without considering the results of the observation.
- $\Pr(D|H, I)$ is the sampling distribution which tells us the probability assigned to the particular data D given the hypothesis H and the background information I.
- $\Pr(H|D, I)$ is the posterior probability of the hypothesis H, considering both the background information I and the result D of the observation.
- $\Pr(D|I)$ is the total probability of the observation being made given the background information. We'll have a lot more to say about this later, but for example if H_1, H_2, ... are a set of mutually exclusive hypotheses, it can be constructed as $\Pr(D|I) = \Pr(D|H_1, I)\Pr(H_1|I) + \Pr(D|H_2, I)\Pr(H_2|I) + \cdots$.

Note that if you conduct two observations which result in data D_1 and D_2, probability theory gives an easy way to combine the data. If we let $D = D_2, D_1$ in (1.8) above, we get

$$\Pr(H|D_2, D_1, I) = \frac{\Pr(D_2, D_1|H, I)\Pr(H|I)}{\Pr(D_2, D_1|I)} \tag{1.9}$$

On the other hand, the product rule of probabilities tells us

$$\Pr(D_2, D_1|H, I) = \Pr(D_2|D_1, H, I)\Pr(D_1|H, I) \tag{1.10a}$$

and

$$\Pr(D_2, D_1|I) = \Pr(D_2|D_1, I)\Pr(D_1|I) \tag{1.10b}$$

[4] See McElreath Chapter 3: "Sampling the Imaginary" for a lot more on this point.

so

$$\begin{aligned}\Pr(H|D_2, D_1, I) &= \frac{\Pr(D_2|D_1, H, I)\Pr(D_1|H, I)\Pr(H|I)}{\Pr(D_2|D_1, I)\Pr(D_1|I)} \\ &= \frac{\Pr(D_2|D_1, H, I)\Pr(H|D_1, I)}{\Pr(D_2|D_1, I)}\end{aligned} \tag{1.11}$$

where we have applied Bayes's theorem to the first experiment to identify $\Pr(H|D_1, I)$ in the last step. But now we see that

$$\Pr(H|D_2, D_1, I) = \frac{\Pr(D_2|H, D_1, I)\Pr(H|D_1, I)}{\Pr(D_2|D_1, I)} \tag{1.12}$$

is just Bayes's theorem applied to the second experiment, using the posterior of the first experiment $\Pr(H|D_1, I)$, as the prior of the second experiment. So we get the same answer whether we analyze all the results $D = D_2, D_1$ at once, or analyze D_1 first and then update our probability using D_2.

Thursday 26 January 2017

2 Estimation of a Single Parameter

2.1 Notation

Most of the propositions we will be interested in will concern the values of parameters, so rather than a probability $\Pr(A|I)$ for a proposition A, we'll consider a probability distribution $p(\theta|I)$ for a parameter θ. In the case of a discrete parameter, this will be what you know as the probability mass function

$$p(\theta|I) = \Pr(\Theta = \theta|I) \tag{2.1}$$

Note, however, that the formal distinction between a "random variable" Θ and its value θ is one we won't generally worry about, using θ to refer to both. Another notational departure from most formal statistics books, in which we'll follow Gelman, is that we'll also use the notation $p(\theta|I)$ to refer to the probability *density* function when θ is a continuous parameter. So we'll mean something like this in that case

$$p(\theta|I)\, d\theta = \Pr(\theta \le \Theta < \theta + d\theta) \tag{2.2}$$

Whether we need to sum or integrate over θ to get a probability should be apparent from the meaning of θ and the context.[5]

2.2 Bayesian Parameter Estimation

Our first application of Bayesian inference will be to estimation of a parameter θ in light of data $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. Assume for concreteness that θ is continuous. If $p(\theta|I)$ is the prior probability distribution for θ, which should be normalized so that

$$\int_{-\infty}^{\infty} p(\theta|I)\, d\theta = 1 \tag{2.3}$$

and $p(\mathbf{y}|\theta, I)$ is the sampling distribution for $\mathbf{y}$ given the model implied by I and the parameter value θ, then Bayes's theorem tells us that the posterior distribution for θ should be

$$p(\theta|\mathbf{y}, I) = \frac{p(\mathbf{y}|\theta, I)\, p(\theta|I)}{p(\mathbf{y}|I)} \tag{2.4}$$

Note that the posterior must also be normalized so that

$$\int_{-\infty}^{\infty} p(\theta|\mathbf{y}, I)\, d\theta = 1 \tag{2.5}$$

[5] This sounds a bit confusing, but given that we'll be interested in joint probability distributions where some arguments are continuous and others discrete, it will be simpler in the long run to use $p(\theta|I)$ for a pdf rather than changing the letter to f.

On the other hand, the sampling distribution $p(\mathbf{y}|\theta, I)$, which is also the likelihood function for θ, is not a density in θ, and it must be normalized so that if you sum or integrate over all values of $\mathbf{y}$ (*not* θ), you get 1. The denominator $p(\mathbf{y}|I)$ can be evaluated as

$$p(\mathbf{y}|I) = \int_{-\infty}^{\infty} p(\mathbf{y}, \theta|I)\, d\theta = \int_{-\infty}^{\infty} p(\mathbf{y}|\theta, I)\, p(\theta|I)\, d\theta \tag{2.6}$$

Note that we generally don't need to calculate this separately, because it only depends on the data $\mathbf{y}$, not the parameter θ which we're trying to estimate. So in fact, we can write the posterior distribution for θ as

$$p(\theta|\mathbf{y}, I) \propto p(\mathbf{y}|\theta, I)\, p(\theta|I) \tag{2.7}$$

where the proportionality sign $\propto$ means that

$$p(\theta|\mathbf{y}, I) = N(\mathbf{y})\, p(\mathbf{y}|\theta, I)\, p(\theta|I) \tag{2.8}$$

where $N(\mathbf{y})$ is some unspecified constant which doesn't depend on θ (although it can depend on $\mathbf{y}$). We can calculate the constant if necessary at the end of the problem by enforcing the normalization condition (2.5). Of course, we could substitute in (2.8) and solve for

$$N(\mathbf{y}) = \frac{1}{\int_{-\infty}^{\infty} p(\mathbf{y}|\theta, I)\, p(\theta|I)\, d\theta} = \frac{1}{p(\mathbf{y}|I)} \tag{2.9}$$

so of course everything is consistent.

2.2.1 Example: Bernoulli Trials

To make all this concrete, suppose we have a series of n Bernoulli trials, i.e., experiments where there are only two possible results, which we can write as $y_i = 0$ or $y_i = 1$. Let the information I include the model that these trials are independent, and each has the same unknown probability θ, where $0 \le \theta \le 1$, of the "success" result $y_i = 1$. Then the probability distribution for each y_i will be

$$p(y_i|\theta, I) = \left.\begin{cases} \theta & \text{if } y_i = 1 \\ 1-\theta & \text{if } y_i = 0 \end{cases}\right\} = \theta^{y_i}(1-\theta)^{1-y_i} \tag{2.10}$$

and sampling distribution/likelihood function will be

$$\begin{aligned} p(\mathbf{y}|\theta, I) &= \theta^{y_1}(1-\theta)^{1-y_1}\theta^{y_2}(1-\theta)^{1-y_2}\cdots\theta^{y_n}(1-\theta)^{1-y_n} \\ &= \theta^{y_1+y_2+\cdots+y_n}(1-\theta)^{n-(y_1+y_2+\cdots+y_n)} = \theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \end{aligned} \tag{2.11}$$

The prior $p(\theta|I)$ must be some function of θ which is normalized so that

$$\int_0^1 p(\theta|I)\, d\theta = 1 \tag{2.12}$$

and the posterior will be

$$p(\theta|\mathbf{y}, I) \propto \theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}}\, p(\theta|I) \tag{2.13}$$

Note that while the posterior is explicitly constructed to consider all the data $\mathbf{y}$, it depends only on the total number of successes $y_{\rm tot} = \sum_{i=1}^{n} y_i$, and the total number of trials n. In fact, if the only data we had access to consisted of $y_{\rm tot}$ and n, we would have a sampling distribution which was binomial

$$p(y_{\rm tot}|\theta, n, I) = \frac{n!}{y_{\rm tot}!(n-y_{\rm tot})!}\theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \tag{2.14}$$

and a posterior

$$p(\theta|y_{\rm tot}, n, I) \propto p(y_{\rm tot}|\theta, n, I)\, p(\theta|I) \propto \theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}}\, p(\theta|I) \tag{2.15}$$

where we have dropped the binomial coëfficient $\frac{n!}{y_{\rm tot}!(n-y_{\rm tot})!}$ because it is independent of θ and can be absorbed into the normalization constant. That means the posteriors $p(\theta|y_{\rm tot}, n, I)$ and $p(\theta|\mathbf{y}, I)$ are equal up to a θ-independent factor. But since they must each integrate to 1, they must in fact be equal,

$$p(\theta|y_{\rm tot}, n, I) = p(\theta|\mathbf{y}, I) \tag{2.16}$$

which means we'll draw the same inferences about θ if we know all the data $\mathbf{y}$ or only the summary $y_{\rm tot}$.[6]

Choice of Prior Choosing the appropriate prior distribution $p(\theta|I)$ for the state of knowledge I you're trying to describe is a tricky business. If these Bernoulli trials represent flips of a coin, you might choose a prior distribution that reflects that most coins are nearly fair. If they represent outcomes of repeated games between the same two chess players, you might choose one where θ is relatively unlikely to be close to 1/2 (because any two players are unlikely to be evenly matched). One choice is to assume we know nothing about the Bernoulli trials, and consider any value of θ to be equally likely,[7] which gives a prior distribution

$$p(\theta|I) = 1, \qquad 0 \le \theta \le 1 \tag{2.17}$$

Then the posterior distribution will be

$$p(\theta|\mathbf{y}, I) \propto \theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \tag{2.18}$$

Now, it turns out that there is an exact solution to the integral

$$\int_0^1 \theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}}\, d\theta = \frac{y_{\rm tot}!(n-y_{\rm tot})!}{(n+1)!} \tag{2.19}$$

[6]For those of you who've taken Math Stat, this is a reflection of the fact that $y_{\rm tot}$ is a sufficient statistic for θ, given a value for n.

[7]This is also a non-trivial assumption, as we'll see later.

which means that the normalized posterior probability distribution is

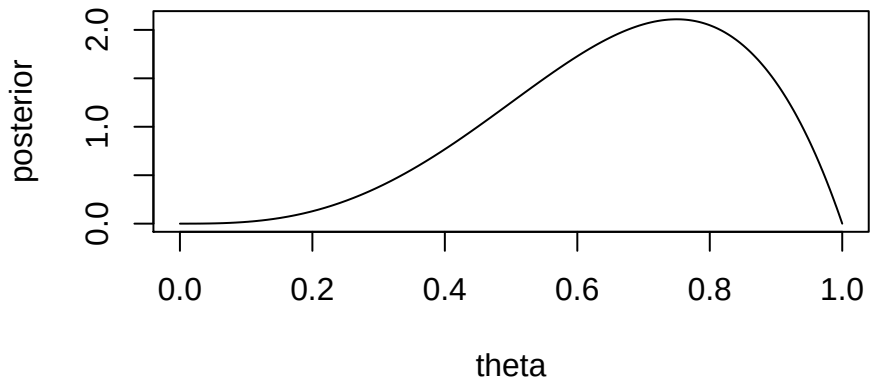
$$p(\theta|\mathbf{y}, I) = \frac{(n+1)!}{y_{\rm tot}!(n-y_{\rm tot})!}\theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \tag{2.20}$$

The posterior distribution in this case is a beta distribution with parameters $\alpha = y_{\rm tot} + 1$ and $\beta = n - y_{\rm tot} + 1$.

Numerical Computation Let's put aside the analytical trick, and consider the problem again from a numerical point of view, using the R commands

```
> theta = seq(from=0, to=1, length.out = 1000)
> d_theta = theta[2]-theta[1]
> prior = dunif(theta)
> ytot = 3
> n = 4
> likelihood = dbinom(ytot,n,theta)
> posterior = prior * likelihood
> posterior = posterior / (sum(posterior) * d_theta)
> plot(theta,posterior,'l')
```

The resulting plot looks like this



We can do various things to summarize this posterior distribution. For example, its expectation value is

$$E(\theta|\mathbf{y}, I) = \int_0^1 \theta\, p(\theta|\mathbf{y}, I)\, d\theta \tag{2.21}$$

Numerically, we can evaluate it in R as

```
> sum(theta*posterior)*d_theta
[1] 0.6666661
```

Note that in general, we can use the properties of the Beta distribution to write

$$E(\theta|\mathbf{y}) = \frac{y_{\text{tot}} + 1}{n + 2} \tag{2.22}$$

which we can verify in this case is $\frac{3+1}{4+2} = \frac{4}{6} = \frac{2}{3}$.

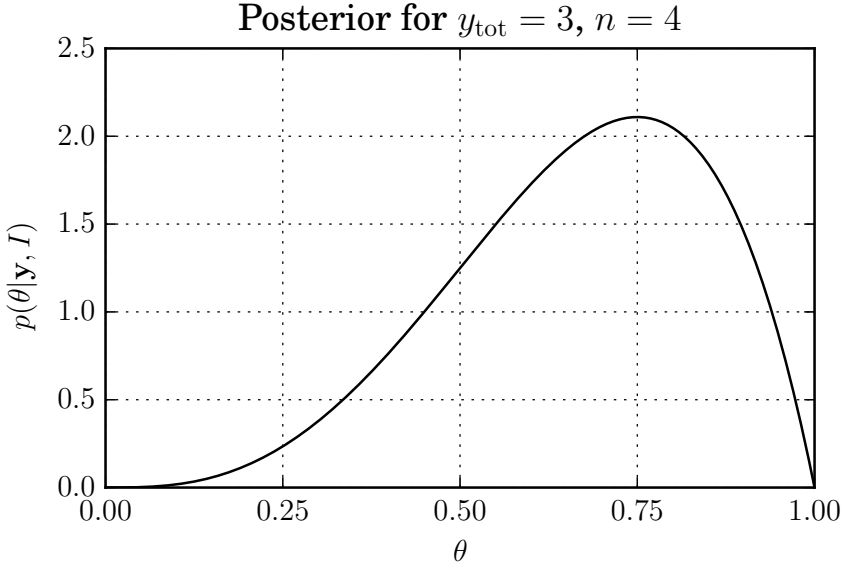
Tuesday 31 January 2017 – Refer to Chapter 2 of Gelman

2.3 Summarizing a Posterior

The result of Bayesian parameter estimation is a posterior probability distribution $p(\theta|\mathbf{y}, I)$. For instance, if we conduct n Bernoulli trials and observe y_{tot} successes, and start with a uniform prior on the probability θ of success, we know the posterior is

$$p(\theta|\mathbf{y}, I) = \frac{(n+1)!}{y_{\text{tot}}!(n - y_{\text{tot}})!}\theta^{y_{\text{tot}}}(1-\theta)^{n-y_{\text{tot}}} \tag{2.23}$$

which looks like this:



Sometimes, though, we're looking for a few numbers to summarize the distribution: what is our best-guess value of θ? How far is θ likely to be away from that value? What are the chances that θ is in a particular range of values? What is a range of values that's reasonably likely to contain θ? Most of these are things you may already know how to construct from a probability distribution. The difference is that now they're direct statements about the parameter of interest, rather than statements about hypothetical alternative data you might have observed.

Here are the definitions and illustrations of some of the quantities you're likely to want to derive from a posterior.

2.3.1 Moments of the posterior distribution.

Expectation Value As with any probability distribution, you can calculate the mean or expectation value of the distribution

to get a point estimate:

$$E(\theta|\mathbf{y}, I) = \int_{-\infty}^{\infty} \theta\, p(\theta|\mathbf{y}, I)\, d\theta \tag{2.24}$$

Variance Likewise, you can get a sense of the spread of the distribution using the variance:

$$\begin{aligned} V(\theta|\mathbf{y}, I) &= E\big([\theta - E(\theta|\mathbf{y}, I)]^2\big) \\ &= \int_{-\infty}^{\infty} [\theta - E(\theta|\mathbf{y}, I)]^2\, p(\theta|\mathbf{y}, I)\, d\theta \end{aligned} \tag{2.25}$$

2.3.2 Percentiles of the posterior distribution.

The posterior is a probability distribution, and so you can use it to make probabilistic statements about the value of θ. In particular,

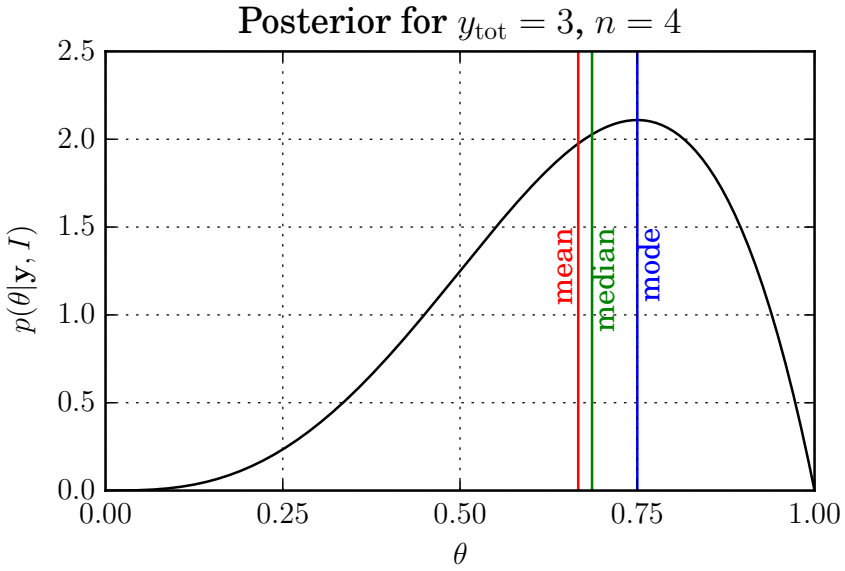
$$\Pr(\theta_1 < \theta < \theta_2|\mathbf{y}, I) = \int_{\theta_1}^{\theta_2} p(\theta|\mathbf{y}, I)\, d\theta \tag{2.26}$$

Sometimes that's what you want to know, like what's the probability that $\theta < \frac{1}{2}$. But more often, to summarize the distribution, you'll want to go the other way, and ask for the theta values which contain a particular fraction of the probability distribution.

Median These probability statements can be used to get a point estimate, the median of the posterior. If we call it $\tilde{\theta}$, it's defined as the value that has half of the posterior probability below and half above.

$$\Pr(\theta < \tilde{\theta}|\mathbf{y}, I) = \int_{-\infty}^{\tilde{\theta}} p(\theta|\mathbf{y}, I)\, d\theta = \frac{1}{2} \tag{2.27}$$

For the posterior we've been considering, we can plot the median, along with the mean:



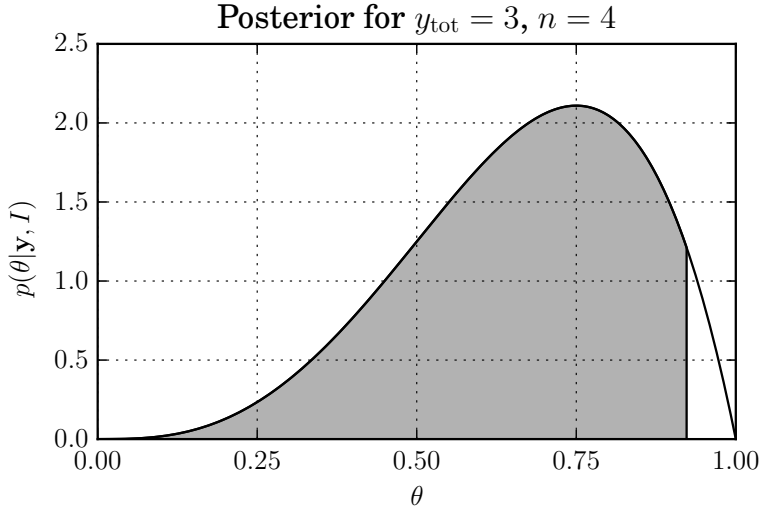
We've also shown the mode, which is the parameter value which maximizes the posterior distribution. A standard name for that is the maximum a posteriori (MAP).

Plausible Intervals More often, we are interested in a range of values which contains some particular probability, say 68% or 90% or 95%, so that

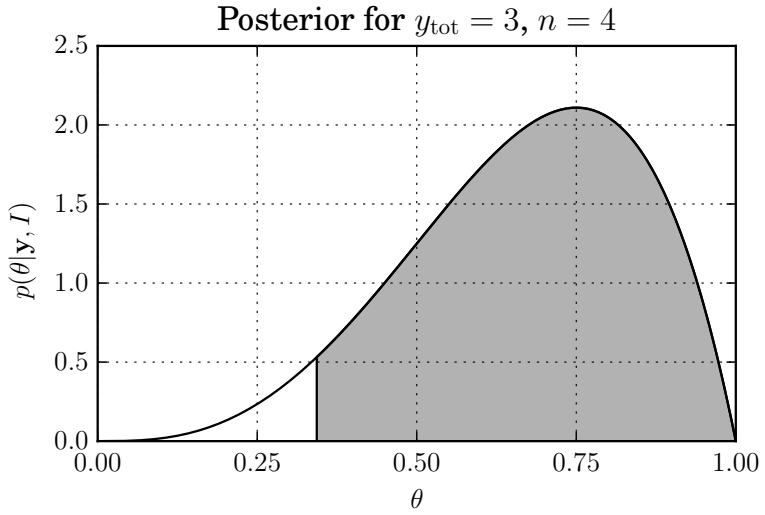
$$\Pr(\theta_\ell < \theta < \theta_u|\mathbf{y}, I) = \int_{\theta_\ell}^{\theta_u} p(\theta|\mathbf{y}, I)\, d\theta = 1 - \alpha \tag{2.28}$$

This is called a plausible interval or credible interval, and it's qualitatively like a confidence interval. But note that, unlike a confidence interval, it really is a statement about the unknown parameter value based on the observation we actually made, and not a statement about the possible outcomes of other random measurements that could have occurred.

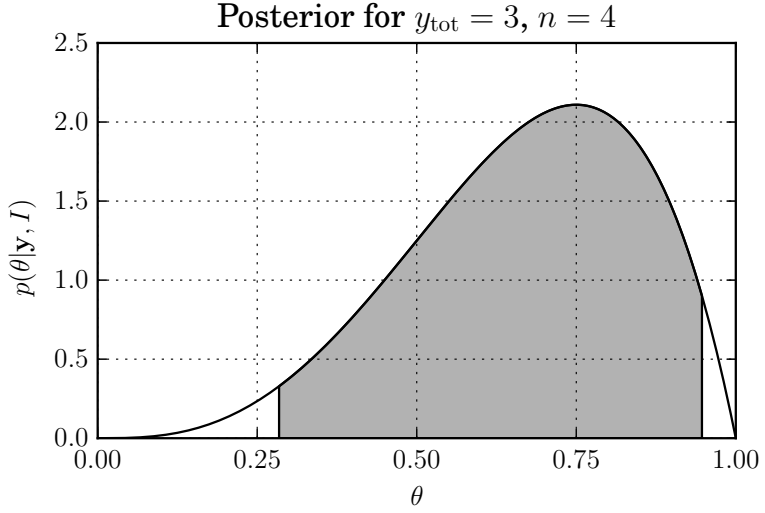
As with a confidence interval, there are many choices of a range of values which include a given total probability. They include an **Upper Limit**, for which $\Pr(\theta < \theta_u) = 1 - \alpha$. We show this for $\alpha = 0.05$, i.e., a 95% upper limit:



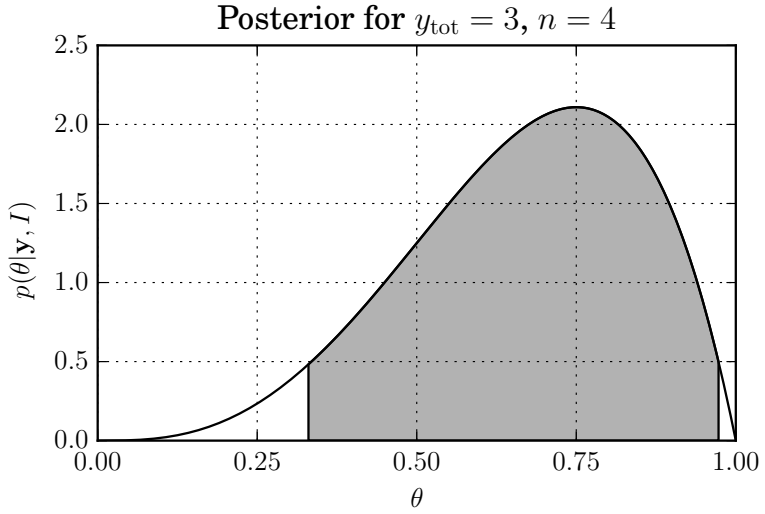
a **Lower Limit**, for which $\Pr(\theta_\ell < \theta) = 1 - \alpha$:



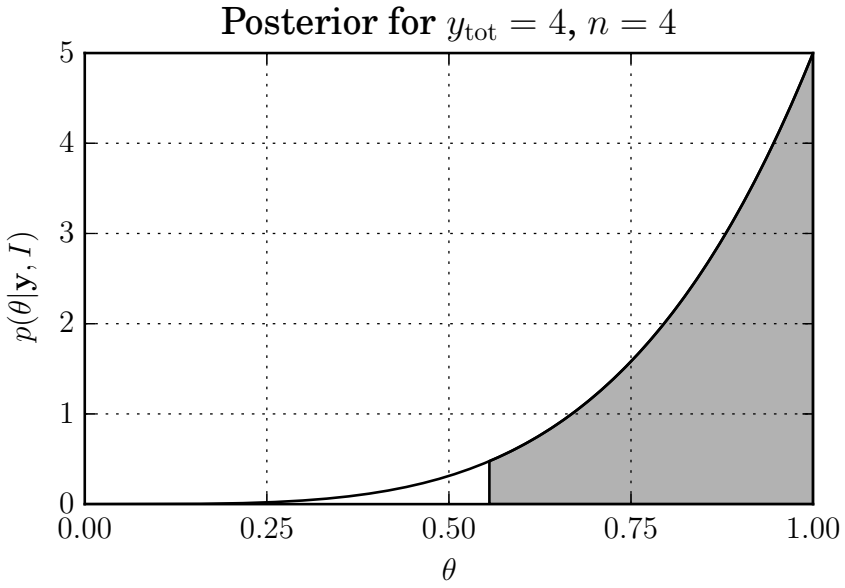
a **Symmetric Interval**, for which $\Pr(\theta < \theta_\ell) = \frac{\alpha}{2} = \Pr(\theta > \theta_u)$



One appealing choice is the **Highest Density Region (HDR)**, which is the plausible interval made up of the parameter values with the highest posterior pdf values (so the MAP value will always be part of the HDR):



If the posterior pdf is continuous, the boundaries of the HDR will be at the same posterior value. One of the appealing features of the HDR is that it can produce either a one-sided or two-sided interval, depending on the shape of the posterior. For instance, if we had seen four successes in four trials, the HDR would produce a lower limit on θ:



2.3.3 Predictive Distributions

Another thing you can do with the posterior distribution for the parameter θ is make predictions about future measurements. Suppose $\tilde{\mathbf{y}}$ is the outcome of a future observation which is also determined by the parameter θ, and has sampling distribution $p(\tilde{\mathbf{y}}|\theta, I)$.[8] It is then reasonable to ask about $p(\tilde{\mathbf{y}}|\mathbf{y}, I)$, the probability distribution you'd assign to the future measurements if you knew about the past measurements. We can get this by constructing the joint probability distribution[9]

$$p(\tilde{\mathbf{y}}, \theta|\mathbf{y}, I) = p(\tilde{\mathbf{y}}|\theta, \mathbf{y}, I)\, p(\theta|\mathbf{y}, I) = p(\tilde{\mathbf{y}}|\theta, I)\, p(\theta|\mathbf{y}, I) \quad (2.29)$$

[8] This may or may not have the same functional form as $p(\mathbf{y}|\theta, I)$. For instance, in our Bernoulli trials experiment, we may be doing an additional $\tilde{n}$ trials, where $\tilde{n}$ need not be the same as n.

[9] The fact that $p(\tilde{\mathbf{y}}|\theta, \mathbf{y}, I) = p(\tilde{\mathbf{y}}|\theta, I)$ is telling us that if we know the actual parameter value θ, knowing the original data $\mathbf{y}$ doesn't tell us any more about the future data $\tilde{\mathbf{y}}$.

and marginalizing

$$p(\tilde{\mathbf{y}}|\mathbf{y}, I) = \int_{-\infty}^{\infty} p(\tilde{\mathbf{y}}, \theta|\mathbf{y}, I)\, d\theta = \int_{-\infty}^{\infty} p(\tilde{\mathbf{y}}|\theta, I)\, p(\theta|\mathbf{y}, I)\, d\theta \tag{2.30}$$

It's worth noting that there are two sources of uncertainty in this posterior predictive distribution. Even if we knew θ, the sampling distribution $p(\tilde{\mathbf{y}}|\theta, I)$ would only make a probabilistic statement about the future data $\tilde{\mathbf{y}}$. But on top of that, we have uncertainty built into the posterior distribution $p(\theta|\mathbf{y}, I)$. (This is why it's important not to just set θ to the most likely value in light of $\mathbf{y}$ and then use that exact value to make future predictions. It understates the uncertainty.)

Of course, we can also apply this technique using the prior distribution rather than the posterior, and obtain what's called a prior predictive distribution

$$p(\mathbf{y}|I) = \int_{-\infty}^{\infty} p(\mathbf{y}, \theta|I)\, d\theta = \int_{-\infty}^{\infty} p(\mathbf{y}|\theta, I)\, p(\theta|I)\, d\theta \tag{2.31}$$

This is of course the denominator of Bayes's theorem, which we've known by other names, such as the marginalized sampling distribution.

2.4 Calculational Techniques

So how do we actually work with posteriors and turn these concepts into quantitative results. There are a few techniques (nicely summarized in *Statistical Rethinking*, BTW.

- **Exact expressions**. If you're lucky, and your problem is particularly simple, you may be able to solve for certain quantities analytically. This is pretty rare in real-world situations (and not so important now that computational power makes numerical techniques feasible), but it is sometimes useful for getting a broad picture of a phenomenon.
- **Grid approximation**. This is the simplest numerical technique, and the one we've used so far. You just lay down a bunch of discrete points in the parameter region of interest, evaluate (or calculate) the posterior at those points, and replace any integrals with sums over the grid. The grid approximation is nice for getting an intuitive sense of what's being calculated, but it doesn't scale very well when you have a lot of parameters to worry about. I can lay down a thousand points in one dimension, but if I have, say, 9 parameters of interest, even 10 points in each direction will give me a 10^9 = one billion-point grid.
- **Gaussian approximation**. If the posterior is simple enough, we can learn a lot about it by approximating it with a Gaussian around the MAP value. More on this in a moment.
- **Monte Carlo/sampling methods**. These will be the topic of the latter part of the course; rather than trying to evaluate the pdf, there are methods that help you effectively draw random numbers whose probability distribution is the posterior pdf of interest.

2.4.1 Gaussian Approximation

One convenient approximation is to expand the posterior about the MAP point, which we'll call $\widehat{\theta}$.[10] We want to use the Taylor expansion

$$f(\theta) = f(\widehat{\theta}) + f'(\widehat{\theta})(\theta - \widehat{\theta}) + \frac{1}{2} f''(\widehat{\theta})(\theta - \widehat{\theta})^2 + \cdots \tag{2.32}$$

But the function we want to expand is not $p(\theta|\mathbf{y}, I)$. That would cause problems, since e.g., if we truncated the expansion at the

[10] Note that we ought to call this $\widehat{\theta}(\mathbf{y})$ or even $\widehat{\theta}(\mathbf{y}, I)$ since it's a property of the posterior $p(\theta|\mathbf{y}, I)$

quadratic term, the resulting parabola would go to negative values, which we know is impossible for a probability distribution. So instead we expand $\ln p(\theta|\mathbf{y}, I)$, to obtain an expansion of the form

$$\begin{aligned}\ln p(\theta|\mathbf{y}, I) = &\ln p(\widehat{\theta}|\mathbf{y}, I) + \left.\frac{\partial \ln p(\theta|\mathbf{y}, I)}{\partial \theta}\right|_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) \\ &+ \frac{1}{2} \left.\frac{\partial^2 \ln p(\theta|\mathbf{y}, I)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} (\theta - \widehat{\theta})^2 + \cdots\end{aligned} \tag{2.33}$$

But the fact that $\widehat{\theta}$ is a local maximum of the function $p(\theta|\mathbf{y}, I)$ (and therefore also of $\ln p(\theta|\mathbf{y}, I)$) tells us two things:

$$\left.\frac{\partial \ln p(\theta|\mathbf{y}, I)}{\partial \theta}\right|_{\theta=\widehat{\theta}} = 0 \tag{2.34a}$$

$$-H(\mathbf{y}, I) = \left.\frac{\partial^2 \ln p(\theta|\mathbf{y}, I)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} < 0 \tag{2.34b}$$

Thus, if we expand to the first non-trivial order, we get

$$\ln p(\theta|\mathbf{y}, I) \approx \ln p(\widehat{\theta}|\mathbf{y}, I) - \frac{1}{2} H(\mathbf{y}, I)(\theta - \widehat{\theta})^2 \tag{2.35}$$

or

$$p(\theta|\mathbf{y}, I) \approx p(\widehat{\theta}|\mathbf{y}, I) e^{-\frac{H}{2}(\theta-\widehat{\theta})^2} \tag{2.36}$$

So we can approximate the posterior, at least near its maximum, with a Gaussian (normal) distribution of mean $\widehat{\theta}(\mathbf{y}, I)$ and variance $\frac{1}{H(\mathbf{y},I)}$.

Note that there's some similarity between the second partial derivative

$$H(\mathbf{y}) = -\left.\frac{\partial^2 \ln p(\theta|\mathbf{y}, I)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} \tag{2.37}$$

and the Fisher information

$$I(\theta) = E\left(-\frac{\partial^2 \ln p(\theta|\mathbf{y}, I)}{\partial \theta^2}\middle|\theta, I\right) \tag{2.38}$$

of classical statistics.

In our Bernoulli trial example, we can actually calculate the parameters of the Gaussian approximation analytically:

$$\ln p(\theta|\mathbf{y}, I) = y_{\text{tot}} \ln \theta + (n - y_{\text{tot}}) \ln(1 - \theta) + \text{const} \tag{2.39}$$

so

$$\frac{\partial \ln p(\theta|\mathbf{y}, I)}{\partial \theta} = \frac{y_{\text{tot}}}{\theta} - \frac{n - y_{\text{tot}}}{1 - \theta} \tag{2.40}$$

and

$$\frac{\partial^2 \ln p(\theta|\mathbf{y}, I)}{\partial \theta^2} = -\frac{y_{\text{tot}}}{\theta^2} - \frac{n - y_{\text{tot}}}{(1 - \theta)^2} \tag{2.41}$$

A little bit of algebra shows that $\widehat{\theta} = \frac{y_{\text{tot}}}{n}$ and $1 - \widehat{\theta} = \frac{n - y_{\text{tot}}}{n}$, so

$$H = \frac{n^2}{y_{\text{tot}}} + \frac{n^2}{n - y_{\text{tot}}} = \frac{n^3}{y_{\text{tot}}(n - y_{\text{tot}})} \tag{2.42}$$

the Gaussian approximation estimates the probability distribution as normal with mean y_{tot}/n and standard deviation

$$\sqrt{\frac{y_{\text{tot}}(n - y_{\text{tot}})}{n^3}} \tag{2.43}$$

Thursday 2 February 2017

2.4.2 Conjugate Prior Distribution Families

On the most recent homework, you showed that, if the prior pdf for a Bernoulli experiment happens to be a beta distribution with parameters $\alpha > 0$ and $\beta > 0$:

$$p(\theta|I_{\alpha,\beta}) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \qquad 0 < \theta < 1 \tag{2.44}$$

the posterior pdf will also be a beta distribution, with parameters $\alpha + y_{\rm tot}$ and $\beta + n - y_{\rm tot}$:

$$p(\theta|\mathbf{y}, I_{\alpha,\beta}) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y_{\rm tot})\Gamma(\beta+n-y_{\rm tot})}\theta^{\alpha+y_{\rm tot}-1}(1-\theta)^{\beta+n-y_{\rm tot}-1} \qquad 0<\theta<1 \quad (2.45)$$

I.e., you increment α by the number of successes and β by the number of failures. We say that the family of beta distributions is the conjugate prior family for the binomial distribution which describes this experiment. That doesn't mean that the prior distribution has to be a member of this family, but if it is, it makes it simple to get the posterior. Note in particular that the uniform prior we've been using is a member of this family, with $\alpha = 1 = \beta$:

$$p(\theta|I_{1,1}) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} = 1 \qquad 0<\theta<1 \quad (2.46)$$

Incidentally, this behavior can be used to argue that the uniform prior we started with doesn't actually describe total ignorance, since it doesn't correspond to the minimum possible values of α and β. If we let α and β go to zero, we obtain what's known as the Haldane prior:

$$p(\theta|I_{0,0}) \propto \frac{1}{\theta(1-\theta)} \qquad 0<\theta<1 \quad (2.47)$$

Note that we can't actually write down the normalization constant because the integral

$$\int_0^1 \frac{d\theta}{\theta(1-\theta)} \quad (2.48)$$

diverges at both endpoints.[11] This is our first example of an *improper prior*. It's not actually a normalized (or normalizable) prior probability distribution, but it's the limit of a family of normalizable distributions. Even though this prior is not normalized, if you go ahead and construct the posterior you get

$$p(\theta|\mathbf{y}, I_{0,0}) \propto \theta^{y_{\rm tot}-1}(1-\theta)^{n-y_{\rm tot}-1} \qquad 0<\theta<1 \quad (2.49)$$

Now, if $0 < y_{\rm tot} < n$, i.e., $\mathbf{y}$ includes at least one success and one failure, the posterior is normalizable, even if the prior wasn't:

$$p(\theta|\mathbf{y}, I_{0,0}) = \frac{\Gamma(n)}{\Gamma(y_{\rm tot})\Gamma(n-y_{\rm tot})}\theta^{y_{\rm tot}-1}(1-\theta)^{n-y_{\rm tot}-1} \qquad 0<\theta<1 \quad (2.50)$$

Working with improper priors is often a convenient shortcut, but if things get confusing, you can always go back to considering them the limiting results of a family of proper priors. In this case, for example, we can find the posterior for small but finite α and β and then take the limit of the posterior as α and β go to zero, and verify that the limit is well defined.

2.5 Reparametrization

There's another way to motivate the improper Haldane prior. Let's define a new parameter

$$\lambda = \ln\frac{\theta}{1-\theta} \quad (2.51)$$

which is the logit or inverse logistic transformation of θ. Since $\frac{\theta}{1-\theta}$ is the odds ratio, the probability of success divided by the probability of failure, λ is the log-odds-ratio associated with the Bernoulli trial. It's not to hard to see that λ is a monotonically increasing function of θ for $0 < \theta < 1$. In the limit that θ goes to 0 from above, λ goes to $-\infty$, and in the limit that θ goes to 1 from below, λ goes to ∞.

Now, we could consider how a probability distribution $p(\theta|I)$ for θ can be converted into the corresponding distribution $p(\lambda|I)$

[11] Put another way, $\Gamma(0)$ is infinite, so the normalization constant $\frac{\Gamma(1)}{\Gamma(0)\Gamma(0)}$ would be zero.

for λ. The key is that θ and λ are continuous parameters, so the probability distributions are probability densities. To convert, we have to require that the probability contained in an infinitesimal range $d\theta$ equals that contained in the corresponding range $d\lambda$:

$$p(\theta|I)\,d\theta = p(\lambda|I)\,d\lambda \tag{2.52}$$

which means

$$p(\lambda|I) = \frac{p(\theta|I)}{d\lambda/d\theta} \tag{2.53}$$

we really can get the ratio of infinitesimal interval widths by just differentiating the transformation (2.51):

$$\frac{d\lambda}{d\theta} = \frac{d}{d\theta}\left(\ln\theta - \ln(1-\theta)\right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1-\theta+\theta}{\theta(1-\theta)} = \frac{1}{\theta(1-\theta)} \tag{2.54}$$

thus

$$p(\lambda|I) = \theta(1-\theta)p(\theta|I) \tag{2.55}$$

of course, to write this as a function of λ, we have to invert the transformation, which gives the logistic transformation

$$\theta = \frac{1}{1+e^{-\lambda}} \qquad \text{so} \qquad 1-\theta = \frac{e^{-\lambda}}{1+e^{-\lambda}} = \frac{1}{1+e^{\lambda}} \tag{2.56}$$

So in particular, in terms of the log-odds-ratio parameter, the Beta function prior becomes

$$p(\lambda|I_{\alpha,\beta}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha}(1-\theta)^{\beta} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1+e^{-\lambda})^{-\alpha}(1+e^{\lambda})^{-\beta} \tag{2.57}$$

So we see that the Haldane prior, which is the limit as α and β go to zero, is uniform in λ:

$$p(\lambda|I_{0,0}) = \text{const} \qquad -\infty < \lambda < \infty \tag{2.58}$$

Among other things, this makes it clear why it's not normalizable; it has a constant value for an infinite range of values. But it also illustrates that it's not so simple to choose a non-informative prior by requiring it to be uniform. The Bayes-Laplace prior is uniform in θ but not in λ

$$p(\theta|I_{1,1}) = 1 \qquad 0 < \theta < 1 \tag{2.59a}$$

$$p(\lambda|I_{1,1}) = (1+e^{-\lambda})^{-1}(1+e^{\lambda})^{-1} \qquad -\infty < \lambda < \infty \tag{2.59b}$$

while the Haldane prior is uniform in λ but not in θ:

$$p(\theta|I_{0,0}) \propto \frac{1}{\theta(1-\theta)} \qquad 0 < \theta < 1 \tag{2.60a}$$

$$p(\lambda|I_{0,0}) = \text{const} \qquad -\infty < \lambda < \infty \tag{2.60b}$$

2.6 Non-informative priors for location and scale parameters

While it's a subtle question what the correct prior is to reflect total ignorance of about the probability parameter for Bernoulli trials, in some cases it's fairly simple, specifically if we're dealing with a location parameter or a scale parameter.

2.6.1 Location Parameter

A model with a sampling probability $p(\mathbf{y}|\theta, I)$ is said to have a location parameter θ if the observed data $\mathbf{y}$ can be written as $y_i = \theta + w_i$ where the probability distribution $p(\mathbf{w}|I)$ doesn't depend on θ. In that case, the non-informative prior $p(\theta|I_0)$ for the location parameter θ should be translationally invariant, i.e.,

$$\begin{aligned}\int_{\theta_1}^{\theta_2} p(\theta|I_0)\,d\theta &= \Pr(\theta_1 < \theta < \theta_2|I_0) \\ &= \Pr(\theta_1 + c < \theta < \theta_2 + c|I_0) = \Pr(\theta_1 < \theta - c < \theta_2|I_0) \\ &= \int_{\theta_1}^{\theta_2} p(\theta - c|I_0)\,d\theta\end{aligned} \tag{2.61}$$

The only way this can be true for all θ_1 and θ_2 is if $p(\theta - c|I_0) = p(\theta|I_0)$, i.e.,

$$p(\theta|I_0) = \text{constant} \tag{2.62}$$

This is an improper prior of course, so really has to be thought of as the limit of some family like

$$p(\theta|I_0) = \frac{1}{2\Theta} \qquad -\Theta < \theta < \Theta \tag{2.63}$$

2.6.2 Scale Parameter

A model with a sampling probability $p(\mathbf{y}|\theta, I)$ is said to have a scale parameter θ if the observed data $\mathbf{y}$ can be written as $y_i = \theta w_i$ where the probability distribution $p(\mathbf{w}|I)$ doesn't depend on θ. Now the non-informative prior $p(\theta|I_0)$ for the scale parameter θ should be scale invariant. The easiest way to do this is to note that

$$\ln y_i = \ln \theta + \ln w_i \tag{2.64}$$

so under the reparametrization $\lambda = \ln \theta$, the scale parameter θ is replaced by a location parameter λ. So the non-informative prior is uniform in λ:

$$p(\lambda|I_0) = \text{constant} \tag{2.65}$$

If we convert the pdf in λ back to a pdf in θ, we get

$$p(\theta|I_0) = p(\lambda|I_0)\frac{d\lambda}{d\theta} = \frac{p(\lambda|I_0)}{\theta} \propto \frac{1}{\theta} \qquad 0 < \theta < \infty \tag{2.66}$$

This is also an improper prior, of course.

2.7 Case study: Gaussian Likelihoods

To expand our sense of how the prior distribution of a parameter is modified by observations to give us the posterior, consider the relatively simple case where the data are modelled as a set of n independent measurements to estimate the parameter θ, with Gaussian errors which have known standard deviations $\{\sigma_i\}$, which makes the likelihood function

$$p(\mathbf{y}|\theta, I) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - \theta)^2\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \sigma_i^{-2}(y_i - \theta)^2\right) \tag{2.67}$$

We can simplify this by repeated application of the mathematical trick of completing the square:

$$a(x-\theta)^2 + b(y-\theta)^2 = (a+b)\left(\theta - \frac{ax+by}{a+b}\right)^2 + \text{constant} \tag{2.68}$$

where by "constant", we mean a term independent of θ. This means that

$$p(\mathbf{y}|\theta, I) \propto \exp\left(-\frac{1}{2\sigma_{\overline{y}}^2}(\theta - \overline{y})^2\right) \tag{2.69}$$

where

$$\overline{y} = \frac{\sum_{i=1}^{n} \sigma_i^{-2} y_i}{\sum_{i=1}^{n} \sigma_i^{-2}} \tag{2.70}$$

is a weighted average of the measurements, with a weighting equal to the inverse variance and

$$\sigma_{\overline{y}}^2 = \frac{1}{\sum_{i=1}^{n} \sigma_i^{-2}} \tag{2.71}$$

In the special case that all of the measurements have the same uncertainty, $\sigma_i = \sigma$, $\overline{y}$ reduces to the sample mean and $\sigma_{\overline{y}}^2 = \sigma^2/n$, which should be a familiar result from introductory classical statistics. The fact that the likelihood (2.69) depends on the data only through the weighted average $\overline{y}$ shows that this is a sufficient statistic for the parameter θ.

If we further assume that the prior on θ is Gaussian

$$p(\theta|I) \propto \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta-\mu_\theta)^2\right) \tag{2.72}$$

we can complete the square again to get the posterior on θ:

$$p(\theta|\mathbf{y},I) \propto p(\mathbf{y}|\theta,I)p(\theta|I) \propto \exp\left(-\frac{1}{2}\left[\sigma_\theta^{-2}(\theta-\mu_\theta)^2+\sigma_{\overline{y}}^{-2}(\theta-\overline{y})^2\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left[(\sigma_\theta^{-2}+\sigma_{\overline{y}}^{-2})\left(\theta-\frac{\sigma_\theta^{-2}\mu_\theta+\sigma_{\overline{y}}^{-2}\overline{y}}{\sigma_\theta^{-2}+\sigma_{\overline{y}}^{-2}}\right)^2\right]\right) \tag{2.73}$$

So we see that the Gaussian distribution is the conjugate prior family for a Gaussian likelihood. If the prior is a Gaussian, the posterior will also be a Gaussian. The mean will be the weighted average of the mean of the posterior and the weighted average of the data, and the variance will be constructed from the variance of the prior and the likelihood.

Tuesday 7 February 2017 – Refer to Chapter 3 of McElreath

2.8 Sampling as a Computational Method

We've mentioned that, given a set of samples from a probability distribution, you can estimate many quantities associated with the distribution. We'll now illustrate how to do this in practice, using the familiar example of Bernoulli trials, with a uniform prior

$$p(\theta|I) = 1 \qquad 0<\theta<1 \tag{2.74}$$

a sampling distribution

$$p(\mathbf{y}|\theta,I) = \theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \tag{2.75}$$

or equivalently the binomial

$$p(y_{\rm tot}|n,\theta,I) = \binom{n}{y_{\rm tot}}\theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \tag{2.76}$$

We've shown that the posterior distribution is a beta distribution

$$p(\theta|\mathbf{y},I) = \frac{\Gamma(n+2)}{\Gamma(y_{\rm tot}+1)\Gamma(n-y_{\rm tot}+1)}\theta^{y_{\rm tot}}(1-\theta)^{n-y_{\rm tot}} \qquad 0<\theta<1 \tag{2.77}$$

We can draw a sample of size N from the posterior distribution, since we know it's a beta distribution:

```
> n = 4
> ytot = 3
> N = 5
> thetasample = rbeta(N, ytot+1, n-ytot+1)
> thetasample
[1] 0.9159046 0.5313724 0.6895402 0.3615219 0.7491285
```

The exact values in this random sample will of course be different for you, and if we generated it again it would also be different for us:

```
> thetasample = rbeta(N, ytot+1, n-ytot+1)
> thetasample
[1] 0.8276773 0.8933439 0.7511602 0.6667999 0.5405257
```

But in fact this is not a random number generator, it's only pseudo-random, so we make sure we can reproduce the exact same set of numbers whenever the code is run by first *seeding* the random number generator:

```
> set.seed(20170207)
> thetasample = rbeta(N, ytot+1, n-ytot+1)
```

```
> thetasample
[1] 0.7642113 0.6134853 0.8822383 0.4125945 0.6273643
> set.seed(20170207)
> thetasample = rbeta(N, ytot+1, n-ytot+1)
> thetasample
[1] 0.7642113 0.6134853 0.8822383 0.4125945 0.6273643
```

A different seed will give different numbers with no apparent connection to ours:

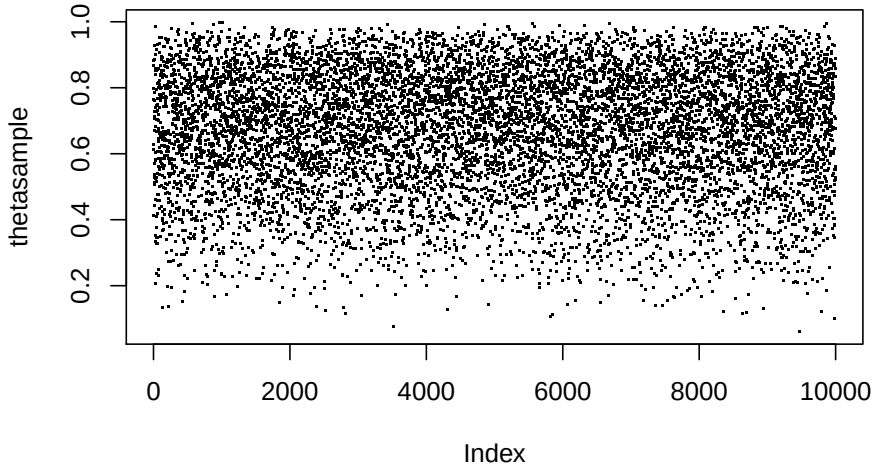
```
> set.seed(20170206)
> thetasample = rbeta(N, ytot+1, n-ytot+1)
> thetasample
[1] 0.9198038 0.8641902 0.4458955 0.7468526 0.1645095
```

Now, let's generate a sample of an interesting size.

```
> set.seed(20170207)
> N = 10000
> thetasample = rbeta(N, ytot+1, n-ytot+1)
```

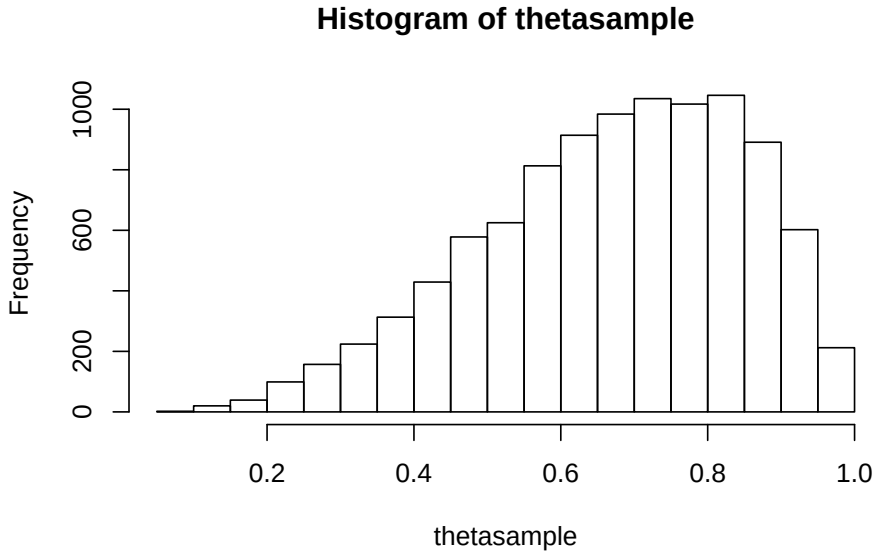
We can plot the sequence of values in the sample:

```
> plot(thetasample,type='p',pch='.')
```



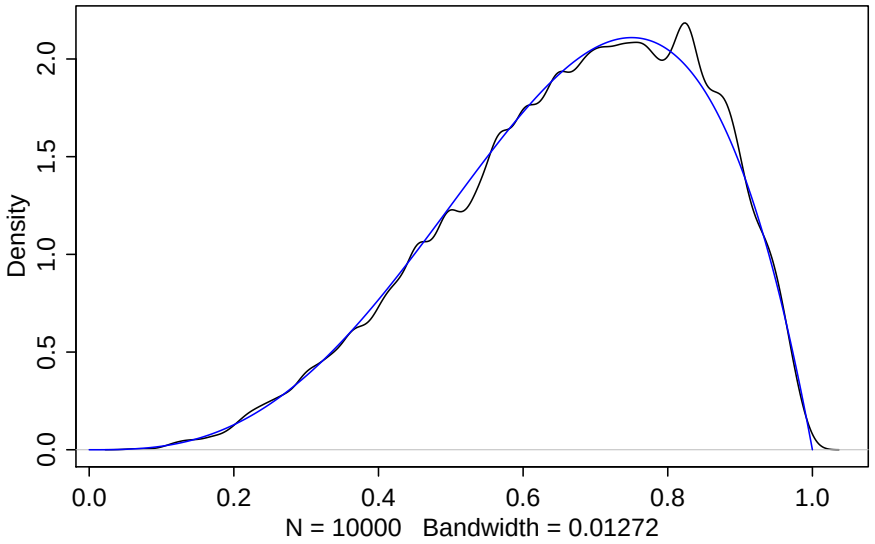
Already we can see that there seem to be more values above 0.5 than below. We can produce a histogram to make this more visible:

```
> hist(thetasample)
```



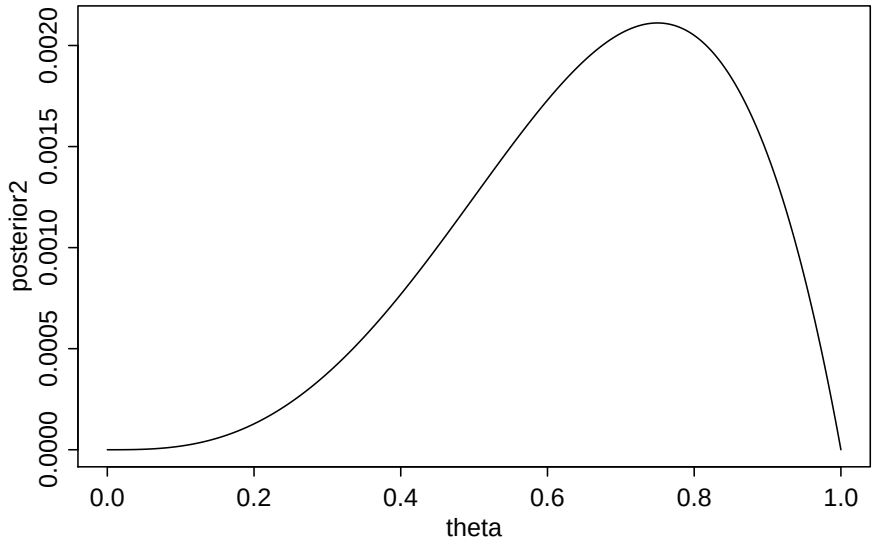
There's the usual question about how to bin the histogram to make it look best, but the rethinking package has a nice function that fits a smooth curve to the sampling histogram and produces an estimate of the posterior density. We can plot this along with the known beta pdf:

```
> library(rethinking)
> dens(thetasample)
> d_theta = 0.001
> theta = seq(from=0, to=1, length.out=1000)
> posterior = dbeta(theta,ytot+1,n-ytot+1)
> lines(theta,posterior,col="blue")
```



Note that in this problem we knew the posterior pdf was a beta distribution, which was useful both for plotting and for generating the sample. But in a more complicated problem that might not be the case. We can instead construct the posterior on a grid using Bayes's theorem and the uniform prior:

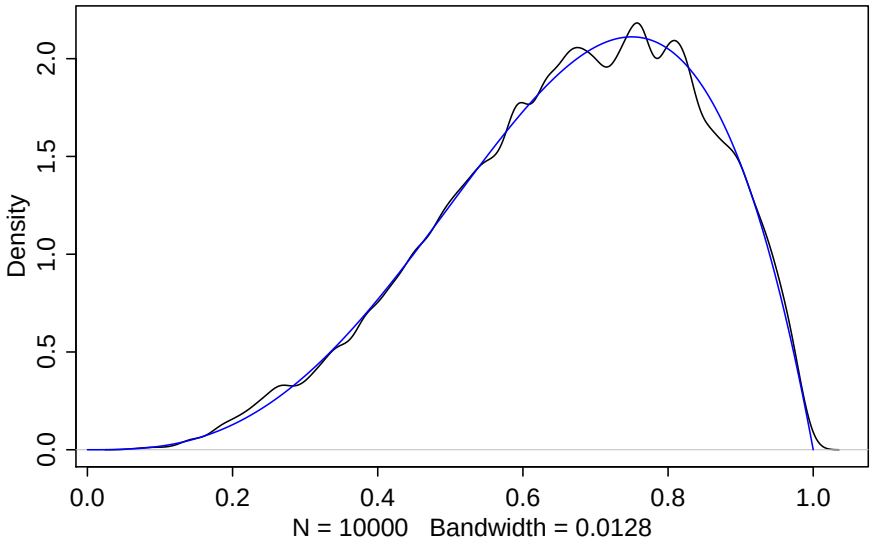
```
> posterior2 = dbinom(ytot,n,theta)
> posterior2 = posterior2/sum(posterior2)
> plot(theta,posterior2,type='l')
```



Note that we've normalized the posterior so that its sum is 1, not its integral. This means we're talking about the posterior probability $p(\theta|\mathbf{y}, I)\, d\theta$ associated with each point on the grid, rather than the probability *density*.

We can also generate the sample a different way, by drawing with replacement from our grid of values, with probability given by the posterior, and verify that we get the same distribution:

```
> thetasample2 = sample(theta,size=N,prob=posterior2,
+                       replace=TRUE)
> dens(thetasample2)
> lines(theta,posterior2/d_theta,col="blue")
```



Note that we needed to divide by $d\theta$ to make the comparison plot, since `dens()` does plot the estimated probability density.

Given either this grid approximation or the sample, we can estimate things like the posterior expectation $E(\theta|\mathbf{y}, I) = \int_0^1 \theta\, p(\theta|\mathbf{y}, I)\, d\theta$:

```
> mean(thetasample2)
[1] 0.6655133
> sum(posterior2*theta)
[1] 0.6666661
```

The normalization that the grid posterior represents $p(\theta|\mathbf{y}, I)\, d\theta$ makes it simpler to estimate expectation values which are associated with integrals. We can also work out other summaries of the posterior in a straightforward way:

```
> median(thetasample2)
[1] 0.6826827
> var(thetasample2)
[1] 0.03219942
> sum(posterior2*(theta-sum(posterior2*theta))^2)
[1] 0.0317459
> sd(thetasample2)
[1] 0.179442
```

If we want to integrate up to get the probability that e.g., $\theta < \frac{1}{2}$, we can take advantage of the fact that inequalities applied to the sample end up being boolean vectors. Thus the average is the fraction for which the value is 1 i.e., the statement is true:

```
> mean(thetasample2<0.5)
[1] 0.1888
```

Similarly, since the grid posterior is normalized to sum to one, the probability of the proposition can be got by summing the probability associated with all the values for which it's true:

```
> sum(posterior2[theta<0.5])
[1] 0.1875001
```

For completeness, we verify that the various quantities agree for the samples constructed in two different ways. Of course, they differ slightly because of the randomness of the sampling:

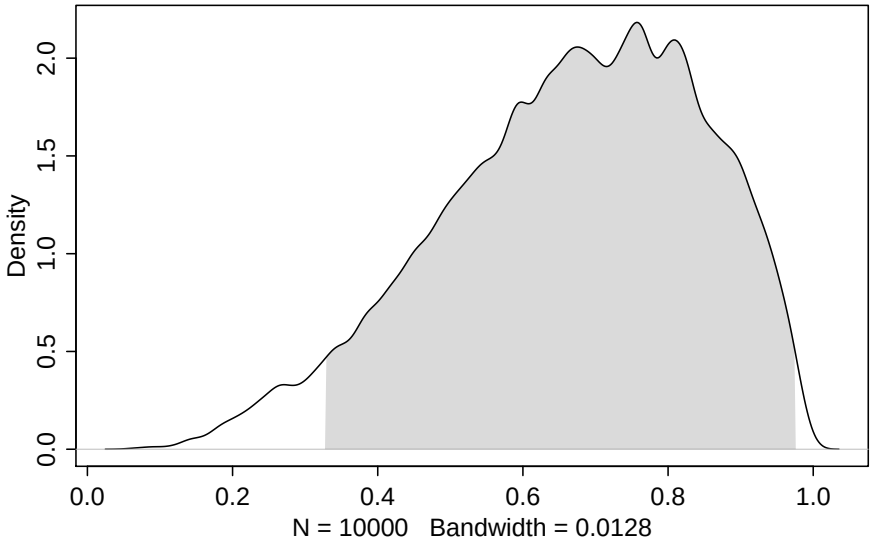
```
> mean(thetasample);mean(thetasample2)
[1] 0.6696085
[1] 0.6655133
> median(thetasample);median(thetasample2)
[1] 0.6902752
[1] 0.6826827
> sd(thetasample);sd(thetasample2)
[1] 0.17831
[1] 0.179442
> mean(thetasample<0.5)
[1] 0.1861
```

If we want to get the highest density interval, `rethinking` happens to have a function for this:

```
> HPDI(thetasample2,prob=0.95)
    |0.95     0.95|
0.3273273 0.9759760
```

Or we can even automatically plot it on the pdf:

```
> dens(thetasample2,show.HPDI = 0.95)
```



One very nice feature of the sampling approach is that it handles reparametrization automatically. If we do the logit transformation $\lambda = \ln \frac{\theta}{1-\theta}$, we showed last time that the posterior pdf is

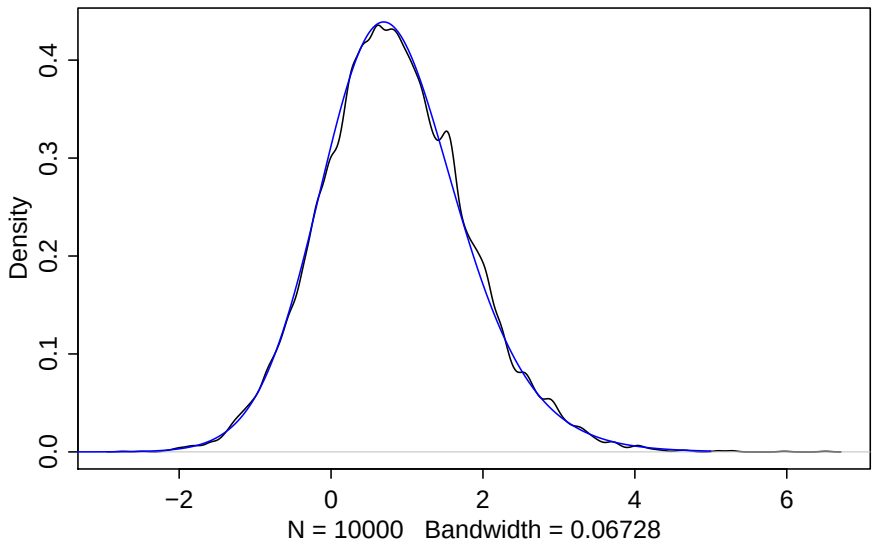
$$p(\lambda|\mathbf{y}, I) = \frac{1}{B(y_{\text{tot}}+1, n-y_{\text{tot}}+1)}(1+e^{-\lambda})^{-(y_{\text{tot}}+1)}(1+e^{\lambda})^{-(n-y_{\text{tot}}+1)} \tag{2.78}$$

where

$$B(y_{\text{tot}}+1, n-y_{\text{tot}}+1) = \frac{\Gamma(y_{\text{tot}}+1)\Gamma(n-y_{\text{tot}}+1)}{\Gamma(n+2)} \tag{2.79}$$

This density conversion comes out automatically if we just apply the transformation to each value of the sample:

```
> lambdasample=logit(thetasample)
> dens(lambdasample)
> lambda3 = seq(from=-5,to=5,length.out=1000)
> pdf_lambda3 = (
+     (1+exp(-lambda3))^(-(ytot+1))
+     * (1+exp(lambda3))^(-(n-ytot+1))
+ ) / ( beta(ytot+1,n-ytot+1) )
> lines(lambda3,pdf_lambda3,col="blue")
```



Thursday 9 February 2017 – Refer to Chapter 3 of Gelman

3 Estimation with Multiple Parameters

So far we have considered procedures to estimate a single parameter θ, but it's often the case that models can contain multiple parameters. Notationally, we can represent the m parameters as a vector $\boldsymbol{\theta} \equiv \theta_1, \theta_2, \ldots, \theta_m$. Then most of the quantities are as before. We can write the sampling distribution for a data vector $\mathbf{y}$ as $p(\mathbf{y}|\boldsymbol{\theta}, I)$, and the prior distribution for the parameters as $p(\boldsymbol{\theta}|I)$. This is a probability density in any continuous parameters within $\boldsymbol{\theta}$. The posterior distribution is

$$p(\boldsymbol{\theta}|\mathbf{y}, I) \propto p(\mathbf{y}|\boldsymbol{\theta}, I) p(\boldsymbol{\theta}|I) \tag{3.1}$$

with the normalization chosen so that the posterior distribution is normalized

$$\int p(\boldsymbol{\theta}|\mathbf{y}, I)\, d^m\theta = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\boldsymbol{\theta}|\mathbf{y}, I)\, d\theta_1 \cdots d\theta_m = 1 \tag{3.2}$$

3.1 Gaussian Approximation and Hessian Matrix

One aspect which behaves slightly differently is expansion about the maximum a posteriori (MAP) estimate of the parameter vector $\boldsymbol{\theta}$. The MAP point is defined by setting all of the partial derivatives of the posterior with respect to the parameters to zero:

$$\left.\frac{\partial \ln p(\boldsymbol{\theta}|\mathbf{y}, I)}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 0 \qquad j = 1, 2, \ldots, m \tag{3.3}$$

We can then do a multi-variable Taylor expansion of the log-likelihood

$$\ln p(\boldsymbol{\theta}|\mathbf{y}, I) = \ln p(\widehat{\boldsymbol{\theta}}|\mathbf{y}, I) + \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{m} \left.\frac{\partial^2 \ln p(\boldsymbol{\theta}|\mathbf{y}, I)}{\partial\theta_j\, \partial\theta_k}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} (\theta_j - \widehat{\theta}_j)(\theta_k - \widehat{\theta}_k) + \cdots \tag{3.4}$$

If we define the Hessian matrix $\mathbf{H}$ by

$$H_{jk} = \left.\frac{\partial^2[-\ln p(\boldsymbol{\theta}|\mathbf{y}, I)]}{\partial\theta_j\, \partial\theta_k}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{3.5}$$

This then gives us a Gaussian approximation

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, I) &\approx p(\widehat{\boldsymbol{\theta}}|\mathbf{y}, I) \exp\left(-\frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{m}(\theta_j - \widehat{\theta}_j) H_{jk} (\theta_k - \widehat{\theta}_k)\right) \\ &\propto \exp\left(-\frac{1}{2}[\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y}, I)]^{\mathrm{T}} \mathbf{H}(\mathbf{y}, I) [\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\mathbf{y}, I)]\right) \end{aligned} \tag{3.6}$$

where e.g., $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_m \end{pmatrix}$ is being treated as a column vector and $\boldsymbol{\theta}^{\mathrm{T}}$ is the row vector which is its transpose. See AppendixA for some reminders about the relevant matrix notation. This approximate distribution is known as a multivariate normal or multivariate Gaussian distribution.

3.2 Marginalization

It is often the case that we care about some parameters and not others. For simplicity, assume $m = 2$, and we're interested in θ_1 but not θ_2. (The extension to multiple parameters in each set

is straightforward.) The marginal posterior distribution for θ_1 will be

$$p(\theta_1|\mathbf{y}, I) = \int_{-\infty}^{\infty} p(\theta_1, \theta_2|\mathbf{y}, I)\, d\theta_2 \tag{3.7}$$

Note that in principle, if all we care about is θ_1, we could already marginalize over θ_2 at the likelihood stage, since

$$p(\theta_1|\mathbf{y}, I) \propto p(\mathbf{y}|\theta_1, I)\, p(\theta_1|I) \tag{3.8}$$

and

$$\begin{aligned} p(\mathbf{y}|\theta_1, I) &= \frac{p(\mathbf{y}, \theta_1|I)}{p(\theta_1|I)} = \frac{\int_{-\infty}^{\infty} p(\mathbf{y}, \theta_1, \theta_2|I)\, d\theta_2}{p(\theta_1|I)} \\ &= \int_{-\infty}^{\infty} p(\mathbf{y}, \theta_2|\theta_1, I)\, d\theta_2 \\ &= \int_{-\infty}^{\infty} p(\mathbf{y}|\theta_1, \theta_2, I)\, p(\theta_2|\theta_1, I)\, d\theta_2 \end{aligned} \tag{3.9}$$

However, we will often be interested in multiple parameters, in which case it's easier to work with the full posterior $p(\boldsymbol{\theta}|\mathbf{y}, I)$.

3.2.1 Marginalization and Parameter Accuracy

To get a sense of the effects of marginalization, suppose that we have a posterior for two parameters which is of the Gaussian form

$$\begin{aligned} p(\theta_1, \theta_2|\mathbf{y}, I) \propto \exp\Bigg(-\frac{1}{2}\,[&H_{11}(\theta_1 - \widehat{\theta}_1)^2 \\ &+ 2H_{12}(\theta_1 - \widehat{\theta}_1)(\theta_2 - \widehat{\theta}_2) + H_{22}(\theta_2 - \widehat{\theta}_2)^2]\Bigg) \end{aligned} \tag{3.10}$$

For concreteness, let's assume $\mathbf{H} = \begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}$ and $\widehat{\boldsymbol{\theta}} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$, and use R to make a contour plot of the log-posterior:

```
> H = matrix(c(3,-2,-2,2),ncol=2)
> theta1hat = 4
> theta2hat = 3
> Ngrid = 101
> theta1 = seq(from=theta1hat-4,to=theta1hat+4,
+              length.out=Ngrid)
> theta2 = seq(from=theta2hat-4,to=theta2hat+4,
+              length.out=Ngrid)
> ones = rep(1,Ngrid)
> theta1grid = outer(theta1,ones,'*')
> theta2grid = outer(ones,theta2,'*')
```

We've generated two 101×101 matrices `theta1grid` and `theta2grid` to act as a grid of θ_1, θ_2 values. Each row of `theta1grid` is filled with the corresponding θ_1 value:

```
> theta1grid[1:5,1:5]
     [,1] [,2] [,3] [,4] [,5]
[1,] 0.00 0.00 0.00 0.00 0.00
[2,] 0.08 0.08 0.08 0.08 0.08
[3,] 0.16 0.16 0.16 0.16 0.16
[4,] 0.24 0.24 0.24 0.24 0.24
[5,] 0.32 0.32 0.32 0.32 0.32
```

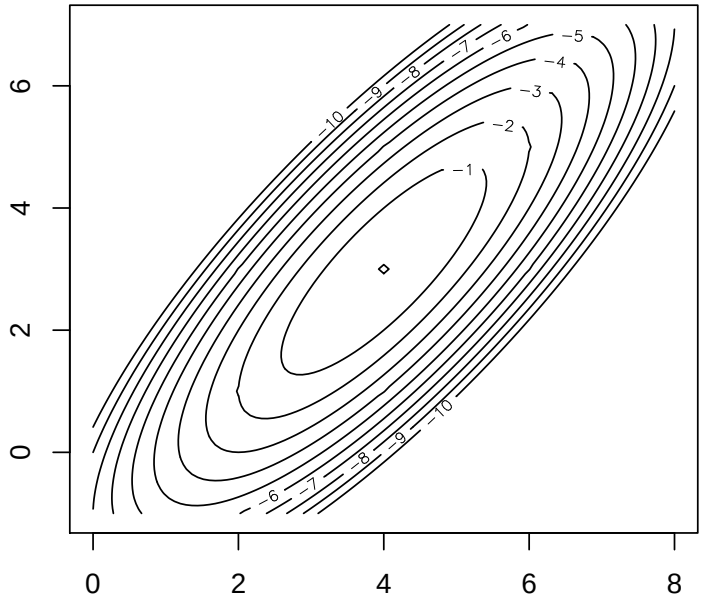
and column of `theta2grid` is filled with the corresponding θ_2 value:

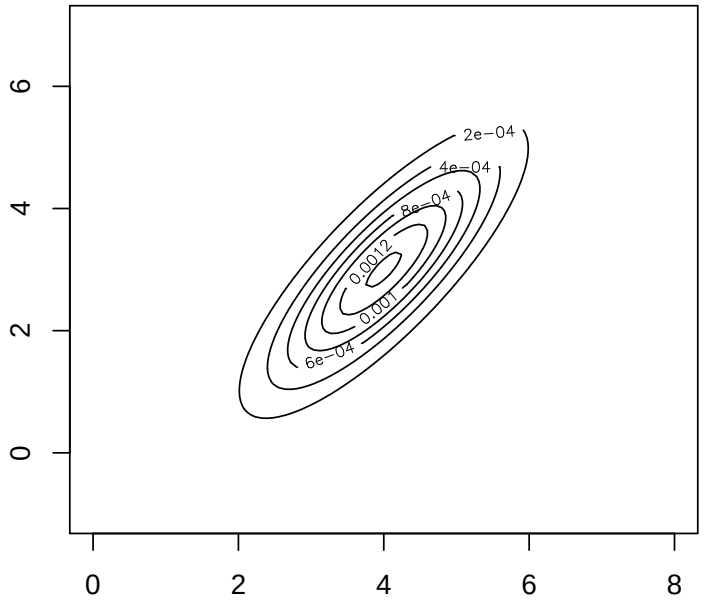
```
> theta2grid[1:5,1:5]
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,]   -1 -0.92 -0.84 -0.76 -0.68
[2,]   -1 -0.92 -0.84 -0.76 -0.68
[3,]   -1 -0.92 -0.84 -0.76 -0.68
[4,]   -1 -0.92 -0.84 -0.76 -0.68
[5,]   -1 -0.92 -0.84 -0.76 -0.68
> logpost = -0.5 * (
```

```
+     H[1,1] * (theta1grid-theta1hat)^2
+     + 2 * H[1,2] * (theta1grid-theta1hat)
+     *(theta2grid-theta2hat)
+     + H[2,2] * (theta2grid-theta2hat)^2
+ )
```

Note that we've defined the log-posterior (which we know up to an additive constant) so that it's maximum value is zero, but that may not always be the case, so it's useful to impose it at this stage, in order to avoid underflow or overflow when we exponentiate it:

```
> max(logpost)
[1] 0
> logpost = logpost - max(logpost)
> contour(theta1,theta2,logpost,levels=0:-10)
> posterior = exp(logpost)
> posterior = posterior / sum(posterior)
> contour(theta1,theta2,posterior)
```





Since we have the posterior evaluated on a grid, we can do the marginalization integral by summing up all the posterior values at a given θ_1:

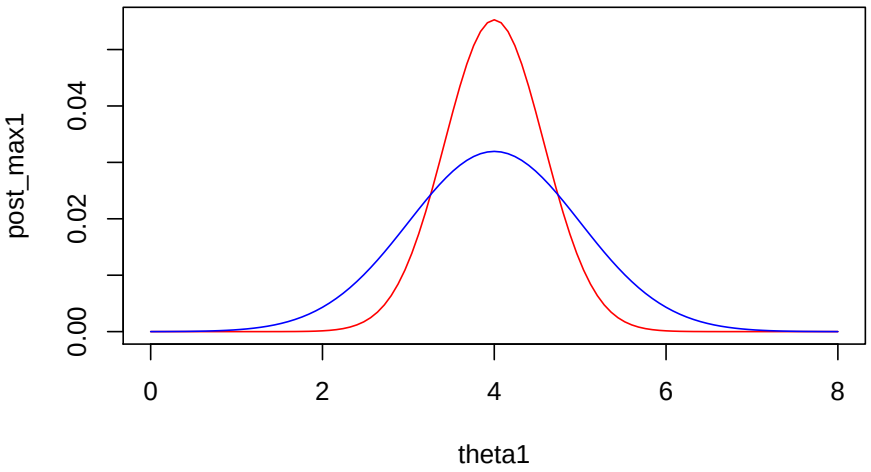
$$p(\theta_1|\mathbf{y}, I)\, d\theta_1 = \int_{-\infty}^{\infty} p(\theta_1, \theta_2|\mathbf{y}, I)\, d\theta_1\, d\theta_2 \tag{3.11}$$

I don't know of a clever way to vectorize this, so I'll just use a loop to sum all of the entries in each row, after initializing a vector of the appropriate length to zero:

```
> post_marg1 = 0.0 * theta1
> for (i in 1:Ngrid){post_marg1[i]=sum(posterior[i,])}
```

Note that this is automatically normalized to sum to 1, the way we've constructed it. Note also that we won't get the right posterior for θ_1 if we insert our best guess value for θ_2 and write something like $p(\theta_1|\widehat{\theta}_2, \mathbf{y}, I)$). Rather than summing the probability surface, this amounts to taking a slice through it at its highest point, and it underestimates the uncertainty in θ_1:

```
> post_max1 = posterior[theta2grid==theta2hat]
> post_max1 = post_max1/sum(post_max1)
> plot(theta1,post_max1,type='l',col='red')
> lines(theta1,post_marg1,col='blue')
```



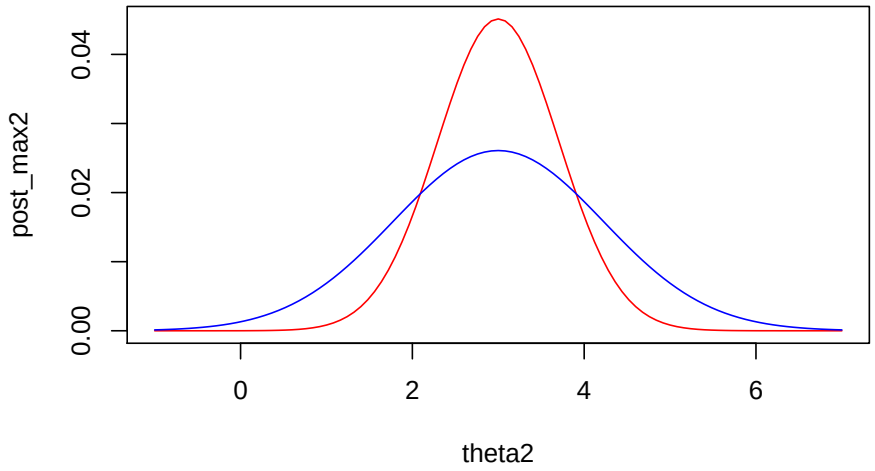
We can also compute the variance of θ_1 using the original grid posterior, or the marginal posterior (in blue above), or the incorrect "slice" distribution (in red above), and see that the first two match, but the third underestimates the variance:

```
> sum(posterior*(theta1grid-theta1hat)^2)
[1] 0.9920522
> sum(post_marg1*(theta1-theta1hat)^2)
[1] 0.9920522
> sum(post_max1*(theta1-theta1hat)^2)
[1] 0.3333333
```

We can also repeat all of this for θ_2 (note that to marginalize, we need to sum all of the entries in each column):

```
> post_marg2 = 0.0 * theta2
> for (i in 1:Ngrid){post_marg2[i]=sum(posterior[,i])}
```

```
> post_max2 = posterior[theta1grid==theta1hat]
> post_max2 = post_max2/sum(post_max2)
> plot(theta2,post_max2,type='l',col='red')
> lines(theta2,post_marg2,col='blue')
```




```
> sum(posterior*(theta2grid-theta2hat)^2)
[1] 1.482658
> sum(post_marg2*(theta2-theta2hat)^2)
[1] 1.482658
> sum(post_max2*(theta2-theta2hat)^2)
[1] 0.4999998
```

Finally, we can estimate the posterior covariance of θ_1 and θ_2 (for which we need the joint posterior):

```
> sum(posterior*(theta1grid-theta1hat)
+     *(theta2grid-theta2hat))
[1] 0.9883074
```

If we put together our numerical estimates for the variances and covariance, it looks like we have a variance-covariance matrix

$$\mathbf{\Sigma} = \text{Cov}(\boldsymbol{\theta}|\mathbf{y}, I) \approx \begin{pmatrix} 1 & 1 \\ 1 & 1.5 \end{pmatrix} \tag{3.12}$$

We can check and see that in fact this is the inverse of the Hessian matrix, $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$:

```
> solve(H)
     [,1] [,2]
[1,]    1  1.0
[2,]    1  1.5
```

This is the general form of the multivariate normal distribution:

$$p(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right) \tag{3.13}$$

In comparison, the "maximum slices" through the posterior gave distributions with variances of $\frac{1}{3} = \frac{1}{H_{11}}$ and $\frac{1}{2} = \frac{1}{H_{22}}$, respectively, which is what we expect from

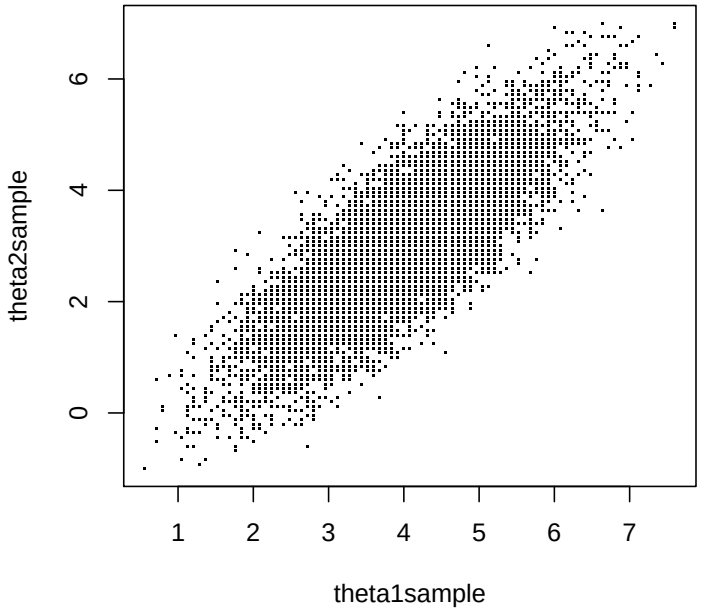
$$p(\theta_1|\widehat{\theta}_2, \mathbf{y}, I) \propto \exp\left(-\frac{1}{2}H_{11}(\theta_1 - \widehat{\theta}_1)^2\right) \tag{3.14a}$$

$$p(\theta_2|\widehat{\theta}_1, \mathbf{y}, I) \propto \exp\left(-\frac{1}{2}H_{22}(\theta_2 - \widehat{\theta}_2)^2\right) \tag{3.14b}$$

Those slices through the posterior miss out on the posterior correlations which are apparent in the full posterior.

We can repeat this calculation using samples from the gridded posterior. In the absence of a clever way to sample from a posterior grid, I'm going to repeatedly draw an index into the grid arrays, taking advantage of the fact that R stores them internally as $10201 = (101)(101)$-element lists.

```
> N = 10000
> set.seed(20170209)
> idxsample = sample.int(n=Ngrid*Ngrid,size=N,
+                        prob=posterior,replace=TRUE)
> theta1sample = theta1grid[idxsample]
> theta2sample = theta2grid[idxsample]
> plot(theta1sample,theta2sample,pch='.')
```

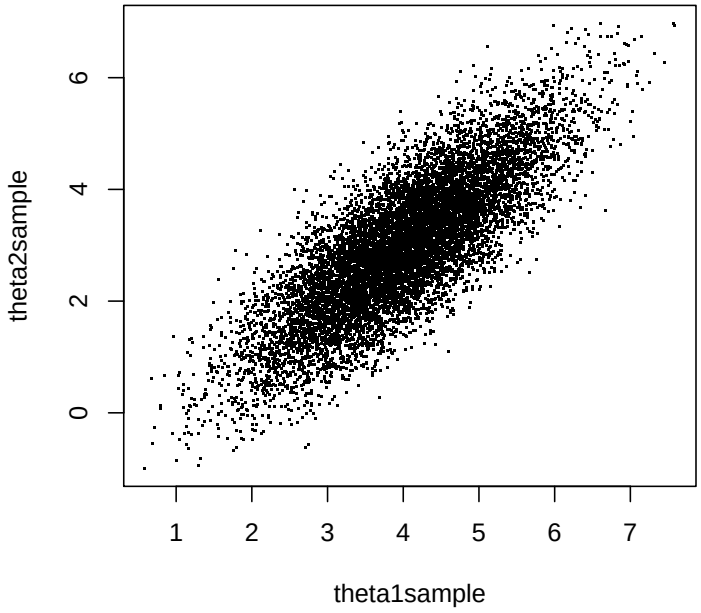


The scatter plot looks a little funny, since it only has values at the grid points, and we can't see if we've chosen the same value multiple times. We can get around this by adding some "jitter" in the form of a random offset in θ_1 uniformly chosen between $-\frac{d\theta_1}{2}$ and $\frac{d\theta_1}{2}$, and likewise for θ_2. In fact, this is more realistic, because we don't believe the posterior is exactly zero away from the grid points; in fact, we hope the probability density is roughly constant in a little box centered on each grid point, or else we didn't choose the grid fine enough.[12]

```
> dtheta1 = theta1[2]-theta1[1]
> dtheta2 = theta2[2]-theta2[1]
> theta1sample = ( theta1grid[idxsample]
+                  + runif(N,min=-0.5*dtheta1,
```

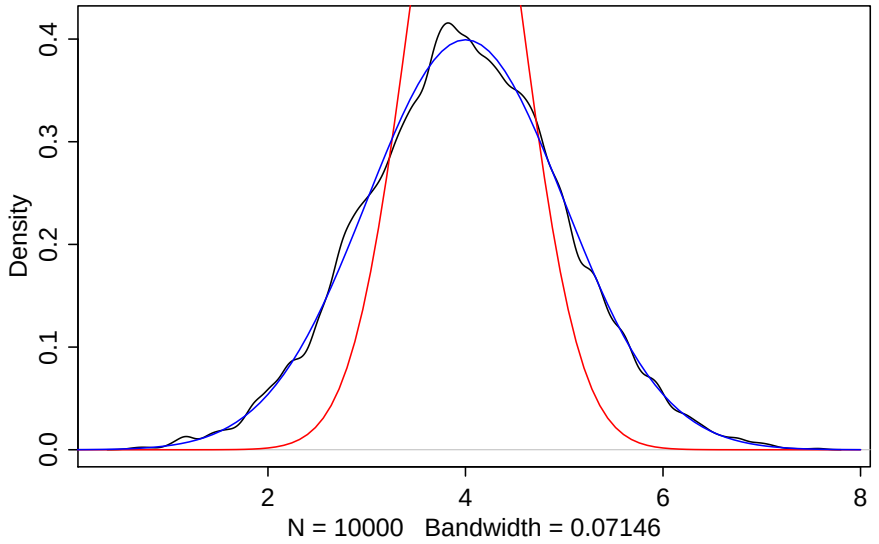
[12]Note that Gelman uses this jitter trick to display the data in Figures 1.1 and 1.2, but in that case it is just for plotting purposes because the variables in question are discrete.

```
+                           max=0.5*dtheta1) )
> theta2sample = ( theta2grid[idxsample]
+                  + runif(N,min=-0.5*dtheta2,
+                           max=0.5*dtheta2) )
> plot(theta1sample,theta2sample,pch='.')
```

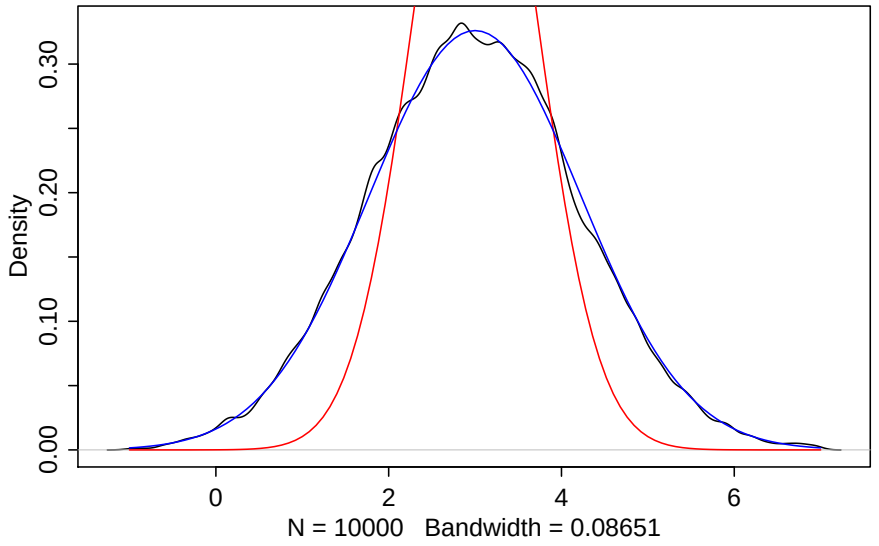


We can work with these samples in exactly the way we did when there was only one parameter, e.g., plotting an estimated pdf:

```
> library(rethinking)
> dens(theta1sample)
> lines(theta1,post_max1/dtheta1,col='red')
> lines(theta1,post_marg1/dtheta1,col='blue')
```




```
> dens(theta2sample)
> lines(theta2,post_max2/dtheta2,col='red')
> lines(theta2,post_marg2/dtheta2,col='blue')
```



and likewise estimate the same means, variances, and covariance

```
> mean(theta1sample)
[1] 3.993434
> mean(theta2sample)
[1] 2.984567
> var(theta1sample)
[1] 1.003869
> var(theta2sample)
[1] 1.471476
> cov(theta1sample,theta2sample)
[1] 0.9870867
```

Tuesday 14 February 2017 – Refer to Chapter 3 of Gelman

3.3 Example: Normal Sample with Unknown Mean and Variance

As an illustration of multiple parameter estimation which leads to familiar results, consider a sample of size n drawn from a normal distribution of unknown mean μ and variance σ^2:

$$p(\mathbf{y}|\mu,\sigma,I) = \prod_{i=1}^{n} p(y_i|\mu,\sigma,I) = \prod_{i=1}^{n} \sqrt{\frac{\sigma^{-2}}{2\pi}} \exp\left(-\frac{1}{2}\sigma^{-2}(y_i-\mu)^2\right)$$
$$= \left(\frac{\sigma^{-2}}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sigma^{-2}(y_i-\mu)^2\right) \tag{3.15}$$

Since μ is a location parameter and σ is a scale parameter, we've argued that the appropriate non-informative prior is uniform in μ and $\ln\sigma$:

$$p(\mu,\sigma|I) \propto \frac{1}{\sigma} \qquad 0<\sigma<\infty \tag{3.16}$$

Note that if we change variables to talk about a density in some power σ^α rather than σ, since $\ln\sigma^\alpha = \alpha\ln\sigma$, such a distribution will also be uniform in $\ln\sigma^\alpha$ as well. So we can just as well write, for instance,

$$p(\mu,\sigma^2|I) \propto \frac{1}{\sigma^2} \qquad 0<\sigma^2<\infty \tag{3.17}$$

or even

$$p(\mu,\sigma^{-2}|I) \propto \frac{1}{\sigma^{-2}} \qquad 0<\sigma^{-2}<\infty \tag{3.18}$$

(Note we have used the fact that $\sigma^{-2}\to\infty$ when $\sigma\to 0$ and $\sigma^{-2}\to 0$ when $\sigma\to\infty$.) Then Bayes's theorem tells us that

$$p(\mu,\sigma^{-2}|\mathbf{y},I) \propto \frac{p(\mathbf{y}|\mu,\sigma^{-2},I)}{\sigma^{-2}} \tag{3.19}$$

with the likelihood

$$p(\mathbf{y}|\mu,\sigma^{-2},I) = p(\mathbf{y}|\mu,\sigma,I) \propto (\sigma^{-2})^{n/2}\exp\left(-\frac{\sigma^{-2}}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \tag{3.20}$$

We've previously shown that completing the square gives

$$\sum_{i=1}^{n}(y_i-\mu)^2 = n(\overline{y}-\mu)^2 + \text{const} \tag{3.21}$$

where $\overline{y}$ is the sample mean $\sum_{i=1}^{n} y_i$. However, that form is not sufficient to attach this problem, since the "constant" (which is independent of the parameters but depends on the data) will get multiplied by σ^{-2} and influence the inference of that quantity. We can work out the constant pretty easily, though, if we write $y_i-\mu = (y_i-\overline{y})-(\mu-\overline{y})$. Then

$$(y_i-\mu)^2 = (y_i-\overline{y})^2 - 2(\mu-\overline{y})(y_i-\overline{y}) + (\mu-\overline{y})^2 \tag{3.22}$$

and

$$\sum_{i=1}^{n}(y_i-\mu)^2 = \sum_{i=1}^{n}(y_i-\overline{y})^2 - 2(\mu-\overline{y})\sum_{i=1}^{n}(y_i-\overline{y}) + n(\mu-\overline{y})^2 \quad (3.23)$$

since $\sum_{i=1}^{n}(y_i-\overline{y})$, the cross term vanishes, and

$$\sum_{i=1}^{n}(y_i-\mu)^2 = (n-1)s^2 + n(\mu-\overline{y})^2 \quad (3.24)$$

where we have used the definition of the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})$. Thus the joint posterior will have the form

$$p(\mu,\sigma^{-2}|\mathbf{y},I) \propto (\sigma^{-2})^{\frac{n}{2}-1}\exp\left(-\frac{\sigma^{-2}}{2}\left[n(\mu-\overline{y})^2+(n-1)s^2\right]\right) \quad (3.25)$$

One thing we can see is that the only relevant information from the data $\mathbf{y}$ are the size n, the sample mean $\overline{y}$ and the sample variance s^2. These three quantities make up the sufficient statistics for the parameters μ and σ^{-2}. It is also instructive to look at the marginal posteriors $p(\mu|\mathbf{y},I)$ and $p(\sigma^{-2}|\mathbf{y},I)$.

3.3.1 Marginal pdf for the mean

First, we marginalize over σ^{-2}:

$$p(\mu|\mathbf{y},I) \propto \int_0^\infty (\sigma^{-2})^{\frac{n}{2}-1}\exp\left(-\frac{\sigma^{-2}}{2}\left[n(\mu-\overline{y})^2+(n-1)s^2\right]\right)d\sigma^{-2} \quad (3.26)$$

Now, this looks like a pretty terrible integral, but if we define

$$\tau = \sigma^{-2}\left[n(\mu-\overline{y})^2+(n-1)s^2\right] \quad (3.27)$$

we end up with

$$p(\mu|\mathbf{y},I) \propto \left[n(\mu-\overline{y})^2+(n-1)s^2\right]^{-n/2}\int_0^\infty \tau^{\frac{n}{2}-1}e^{-\tau/2}\,d\tau \quad (3.28)$$

Now the integral is actually doable (it's a Gamma function), but we don't care what the result is, since it's just a number, independent of μ, which can be absorbed into the normalization constant. If we likewise pull another constant out of what's left, we can write the distribution in the form

$$p(\mu|\mathbf{y},I) \propto \left[1+\frac{1}{n-1}\frac{(\mu-\overline{y})^2}{s^2/n}\right]^{-n/2} \quad (3.29)$$

This shows that the posterior distribution for

$$\frac{\mu-\overline{y}}{\sqrt{s^2/n}} \quad (3.30)$$

is a Student t-distribution with $n-1$ degrees of freedom. So in fact, with this non-informative prior, the plausible interval on μ will be the same as the corresponding frequentist confidence interval!

3.3.2 Marginal pdf for the variance

Now, we marginalize instead over μ:

$$p(\sigma^{-2}|\mathbf{y},I) \propto \int_{-\infty}^{\infty}(\sigma^{-2})^{\frac{n}{2}-1}\exp\left(-\frac{\sigma^{-2}}{2}\left[n(\mu-\overline{y})^2+(n-1)s^2\right]\right)d\mu \quad (3.31)$$

If we define $u = (\mu-\overline{y})\sqrt{\sigma^{-2}}$ we end up with

$$\begin{aligned} p(\sigma^{-2}|\mathbf{y},I) &\propto (\sigma^{-2})^{\frac{n}{2}-1}e^{-(n-1)\sigma^{-2}s^2/2}(\sigma^{-2})^{-1/2}\int_{-\infty}^{\infty}e^{-nu^2/2}\,du \\ &\propto (\sigma^{-2})^{\frac{n-1}{2}-1}\exp\left(-\frac{1}{2}\frac{(n-1)s^2}{\sigma^2}\right) \end{aligned} \quad (3.32)$$

If we stare at this a bit, we'll see it's telling us that the posterior distribution for $\frac{(n-1)s^2}{\sigma^2}$, given n and s, is a chi-squared

with $n-1$ degrees of freedom. Again, the noninformative prior makes the familiar frequentist confidence interval arise as the corresponding Bayesian plausible interval.

If we want to visualize the joint posterior, which will depend on the data only via $\overline{y}$, s^2, and n, it's convenient to change variables from μ and σ^{-2} to

$$t = \frac{\mu - \overline{y}}{\sqrt{s^2/n}} \qquad \text{and} \qquad \chi^2 = (n-1)s^2\sigma^{-2} \tag{3.33}$$

Then the joint posterior becomes

$$p(t, \chi^2|\mathbf{y}, I) \propto (\chi^2)^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}\left[1 + \frac{t^2}{n-1}\right]\chi^2\right) \tag{3.34}$$

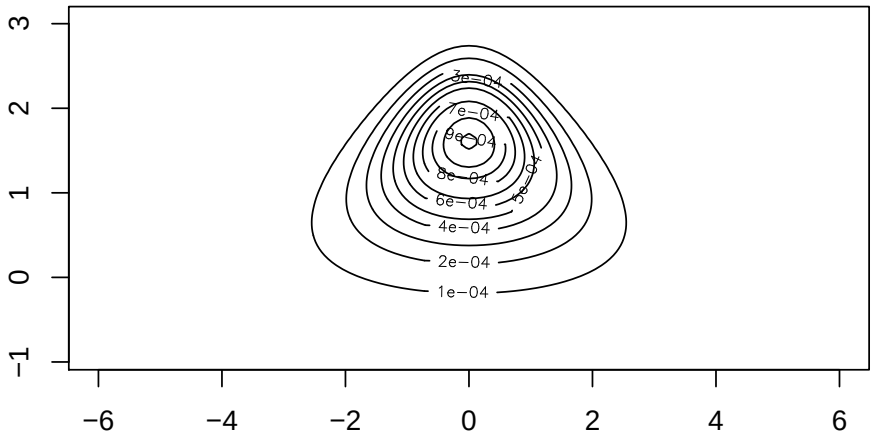
which we see only depends on n, so $\overline{y}$ and s^2 just set the location and scale. To plot this for some choices of n, let's evaluate the pdf on a grid of t and $\ln\chi^2$ values. We'll start with $n = 5$:

```
> Ngrid = 101
> t = seq(from=-6,to=6,length.out=Ngrid)
> ones = rep(1,Ngrid)
> tgrid = outer(t,ones,'*')
> n = 5
> nu = n-1
> chisqmax = nu + 6*sqrt(2*nu)
> chisqmin = nu*(nu/chisqmax)^1.4
```

The range of χ^2 values is kind of arbitrary, but we start with the fact that the expectation value is ν and the standard deviation is 2ν, and we'd like $\ln\nu$ to be close to the middle of the logarithmic range.

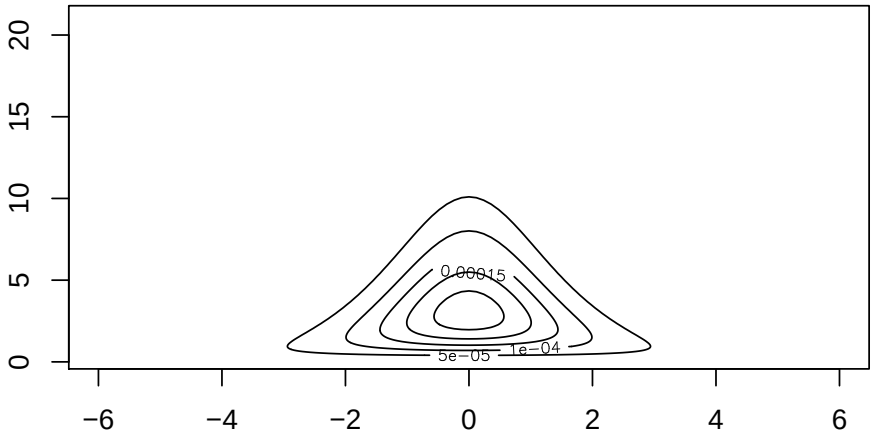
```
> logchisq = seq(from=log(chisqmin),to=log(chisqmax),
+                length.out=Ngrid)
> logchisqgrid = outer(ones,logchisq,'*')
> chisqgrid = exp(logchisqgrid)
> logpost = ( 0.5*n*logchisqgrid
+             - 0.5 * (1+tgrid^2/nu) * chisqgrid
+ )
> max(logpost)
[1] 1.523589
> logpost = logpost - max(logpost)
> posterior = exp(logpost)
> posterior = posterior / sum(posterior)
> contour(t,logchisq,posterior)
```

The contour plot of $p(t, \ln\chi^2|\mathbf{y}, I)$ looks like this



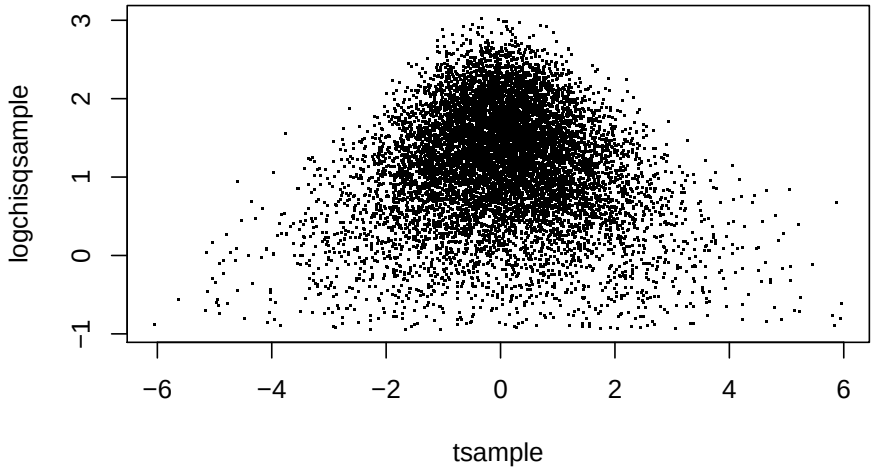
We can also plot $p(t, \chi^2|\mathbf{y}, I) = p(t, \ln\chi^2|\mathbf{y}, I)/\chi^2$:

```
> chisq = exp(logchisq)
> contour(t,chisq,posterior/chisqgrid)
```



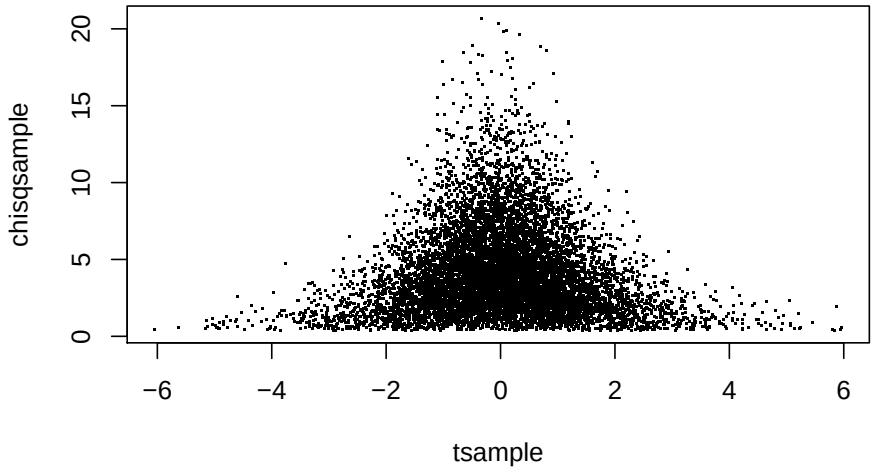
We can sample from the posterior in the usual way:

```
> set.seed(20170214)
> N=10000
> idxsample = sample.int(n=Ngrid*Ngrid,size=N,
+                        prob=posterior,replace=TRUE)
> dt = t[2] - t[1]
> dlogchisq = logchisq[2] - logchisq[1]
> tsample = ( tgrid[idxsample]
+             + runif(N,min=-0.5*dt,
+                     max=0.5*dt) )
> logchisqsample = ( logchisqgrid[idxsample]
+                    + runif(N,min=-0.5*dlogchisq,
+                            max=0.5*dlogchisq) )
> plot(tsample,logchisqsample,pch='.')
```



And finally we can scatter-plot the sampled posterior in t and χ^2:

```
> chisqsample = exp(logchisqsample)
> plot(tsample,chisqsample,pch='.')
```



Sampling from the logarithmic grid enables us to get close to the $\chi^2 = 0$ line while also stretching out to higher χ^2 values.

Thursday 16 February 2017

3.4 Multinomial Distribution

3.4.1 Binomial Distribution Revisited

Another example of a problem with multiple parameters is a multinomial experiment. Recall our previous example of n Bernoulli trials, but let's specialize to the case where we treat the observation as a binomial experiment. Then the sampling distribution for y, the number of successes (which we've called $y_{\rm tot}$ in the past), was

$$p(y|\theta, I) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y}, \qquad y = 0, 1, \ldots, n \quad (3.35)$$

or, thought of as a likelihood function,

$$p(y|\theta, I) \propto \theta^y(1-\theta)^{n-y} \quad (3.36)$$

and the conjugate prior family for the probability θ was the family of beta distributions:

$$p(\theta|I_{\alpha,\beta}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad 0 < \theta < 1 \quad (3.37)$$

We can write this in a way that treats "success" and "failure" more symmetrically, as the first and second of two possible outcomes of the experiment. We then write the number of successes as y_1 and the number of failures as y_2, with the requirement that $y_1 + y_2 = n$. (Note that this is a different notation than what we've used before where $y_i \in \{0, 1\}$ was the outcome of the ith observation. Now y_j, $j \in \{1, 2\}$ is the number of results of the jth kind.) Similarly the probability of success is θ_1 and failure is θ_2, where we require $\theta_1 + \theta_2 = 1$. The sampling distribution or likelihood function is thus

$$p(y_1, y_2|\theta_1, \theta_2, I) = \frac{n!}{y_1!y_2!}\theta_1^{y_1}\theta_2^{y_2}, \qquad y_1, y_2 = 0, 1, \ldots;\ y_1{+}y_2 = n \quad (3.38)$$

or

$$p(y_1, y_2|\theta_1, \theta_2, I) \propto \theta_1^{y_1}\theta_2^{y_2} \quad (3.39)$$

The conjugate prior family is then written

$$p(\theta_1, \theta_2|I_{\alpha_1,\alpha_2}) \propto \theta_1^{\alpha_1}\theta_2^{\alpha_2} \qquad 0 < \theta_1, \theta_2 < 1;\ \theta_1 + \theta_2 = 1 \quad (3.40)$$

The normalization on the probability distribution is a bit funny; it's most easily written using the Dirac delta function, as

$$p(\theta_1, \theta_2|I_{\alpha_1,\alpha_2}) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\delta(\theta_1+\theta_2-1) \qquad 0 < \theta_1, \theta_2 < 1 \quad (3.41)$$

For those of you unfamiliar with the delta function, it's not a function in the strict mathematical sense, but it's operationally defined by

$$\delta(x) = \begin{cases} \infty & x = 0 \\ 0 & x \neq 0 \end{cases} \quad (3.42)$$

and

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1\ ; \quad (3.43)$$

for reasonably well-behaved functions $f(x)$, it will obey

$$\int_{-\infty}^{\infty} f(x)\,\delta(x-a)\, dx = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,\delta(x-a)\, dx = f(a) \quad (3.44)$$

The upshot is that when you integrate the joint pdf,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\theta_1,\theta_2|I_{\alpha_1,\alpha_2})\,d\theta_1\,d\theta_2$$
$$= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1\int_0^1 \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\delta(\theta_1+\theta_2-1)\,d\theta_2\,d\theta_1$$
$$= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 \theta_1^{\alpha_1-1}(1-\theta_1)^{\alpha_2-1}\,d\theta_1 = 1 \quad (3.45)$$

where the delta function means that the $d\theta_2$ integral just sets θ_2 to $1-\theta_1$, and the $d\theta_1$ integral is the Beta function, so the pdf is indeed normalized.

3.4.2 Generalization to Multinomial

To move from the binomial distribution to the multinomial, we suppose that each of the n trials has J possible outcomes rather than just 2. (E.g., we're rolling a 6-sided die rather than flipping a coin.) The data are then $\mathbf{y} \equiv \{y_j\} \equiv y_1, y_2, \ldots, y_J$, the total number of trials which resulted in each of the J outcomes; evidentally $\sum_{j=1}^J y_j = n$. The parameters are $\boldsymbol{\theta} \equiv \{\theta_j\} \equiv \theta_1, \theta_2, \ldots, \theta_J$, the probabilities of each of the outcomes, which must satisfy $\sum_{j=1}^J \theta_j = 1$. The multinomial sampling distribution is

$$p(\mathbf{y}|\boldsymbol{\theta}, I) = \frac{n!}{\prod_{j=1}^J y_j!}\prod_{j=1}^J \theta_j^{y_j} \qquad y_j = 0, 1, \ldots;\ \sum_{j=1}^J y_j = n \quad (3.46)$$

for a likelihood function

$$p(\mathbf{y}|\boldsymbol{\theta}, I) \propto \prod_{j=1}^J \theta_j^{y_j} \quad (3.47)$$

The conjugate prior distribution is a generalization of the beta distribution known as the Dirichlet distribution, which has parameters $\boldsymbol{\alpha} \equiv \{\alpha_j\} \equiv \alpha_1, \alpha_2, \ldots, \alpha_J$:

$$p(\boldsymbol{\theta}|I_{\boldsymbol{\alpha}}) \propto \prod_{j=1}^J \theta_j^{\alpha_j-1} \qquad 0 < \theta_j < 1;\ \sum_{j=1}^J \theta_j = 1 \quad (3.48)$$

The normalization, in case you're curious, is

$$p(\boldsymbol{\theta}|I_{\boldsymbol{\alpha}}) = \frac{\Gamma\left(\sum_{j=1}^J \alpha_j\right)}{\prod_{j=1}^J \Gamma(\alpha_j)}\left(\prod_{j=1}^J \theta_j^{\alpha_j-1}\right)\delta\left(\sum_{j=1}^J \theta_j - 1\right) \quad 0 < \theta_j < 1 \quad (3.49)$$

Note that, unlike $\boldsymbol{\theta}$, which has to sum to 1, and $\mathbf{y}$, which has to sum to n, there's no global requirement for $\boldsymbol{\alpha}$. Each of the elements just has to be positive for normalization purposes.

With this conjugate prior family, the posterior is also a Dirichlet distribution:

$$p(\boldsymbol{\theta}|\mathbf{y}, I) \propto p(\mathbf{y}|\boldsymbol{\theta}, I, p(\boldsymbol{\theta}|I) \propto \prod_{j=1}^J \theta_j^{\alpha_j+y_j-1} \quad (3.50)$$

we've just incremented each α_j by the corresponding y_j.

As in the binomial case, there are different possible choices for a Dirichlet prior. If we set all of the $\alpha_j = 1$, we get a normalized Bayes-Laplace prior which is constant for all of the allowed combinations of the $\{\theta_j\}$. If we take the limit that all of the $\alpha_j \to 0$, we get a Haldane-type improper prior. In that case, the posterior will have $\alpha_j = y_j$, so it will be normalizable if each possible outcome is observed at least once.

Visualizing the joint distribution (prior or posterior) for $\boldsymbol{\theta}$ is a bit tricky. If we consider $0 < \theta_j < 1$, we have a J-dimensional hypercube of unit size. The surface $\sum_{j=1}^J \theta_j = 1$ is then a $J-1$-dimensional simplex connecting the vertices $(1, 0, \ldots, 0)$,

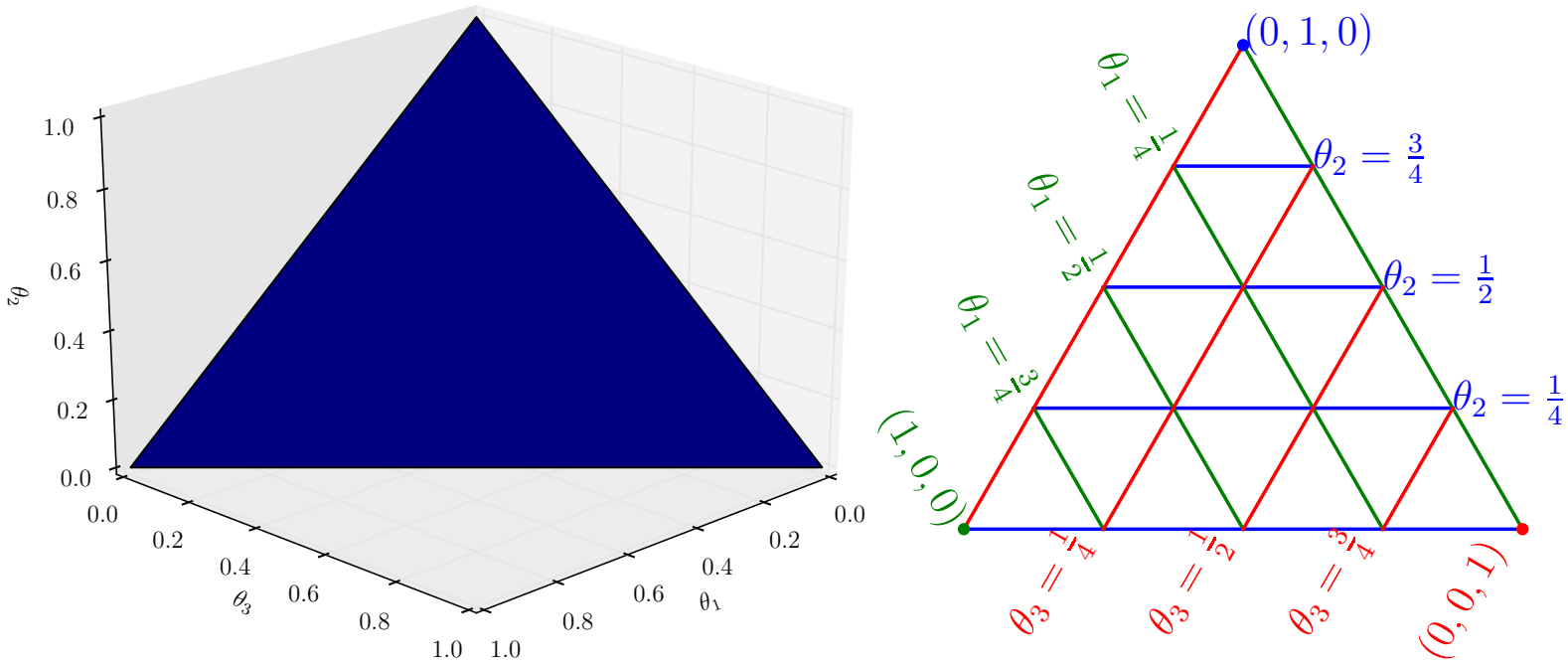


Figure 1: Left: the region of $\theta_1, \theta_2, \theta_3$ space consistent with the constraint $\theta_1 + \theta_2 + \theta_3 = 1$. Right: the grid for a ternary plot on that parameter space.

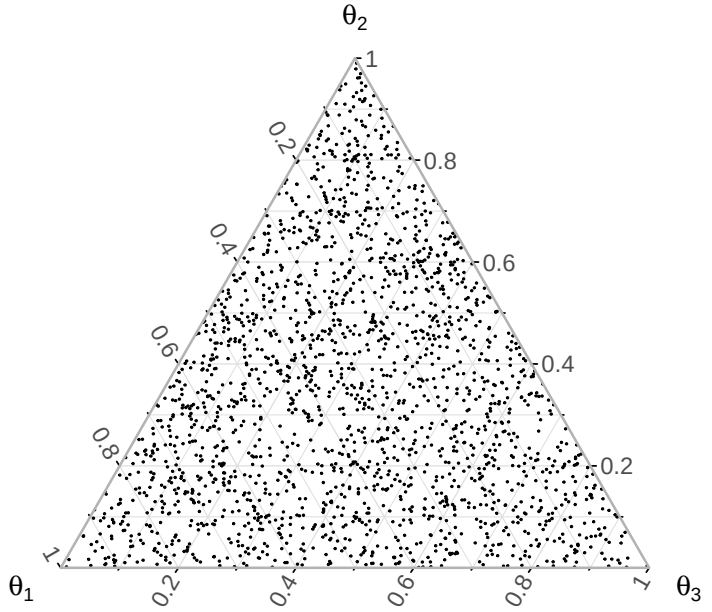
$(0, 1, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 1)$. If this is still too abstract, consider $J = 3$. Then we're talking about a triangle in the corner of the unit cube of $(\theta_1, \theta_2, \theta_3)$, with vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. If we plot things on this triangle, which has three coordinates with a relationship allowing us to specify one given the other two, this is known as a *ternary plot*, as shown in Figure 1 We can make ternary plots of samples drawn from a Dirichlet distribution using the package `ggtern`.[13]

Start with the uniform distribution with $\alpha_1 = \alpha_2 = \alpha_3 = 1$. Note that the Dirichlet distribution is not standard in R (which seems to be lacking in multivariate distributions) so we'll use the `gtools` library:

```
> library(gtools)
> alpha = c(1,1,1)
> N = 2000
> set.seed(20170216)
> thetasample = rdirichlet(N,alpha)
```

Now we make the ternary plot using `ggtern`. Since we'll want to make several of these, we'll define functions that do some of the boilerplate:

```
> library(ggtern)
> myframe = function(mysample) {
+     ( data.frame(theta1=mysample[,1],
+                  theta2=mysample[,2],
+                  theta3=mysample[,3]) )
+ }
> tickvals = seq(.2,1,.2)
> addmyopts = function(myplot) {
+     ( myplot
+       + geom_point(size=0.01)
+       + tern_limits(breaks=tickvals,labels=tickvals)
+       + xlab(expression(theta[1]))
+       + ylab(expression(theta[2]))
+       + zlab(expression(theta[3]))
+       + theme_light()
+     )
+ }
> thetaframe = myframe(rdirichlet(N,alpha))
> addmyopts(ggtern(thetaframe,aes(theta1,theta2,theta3)))
```

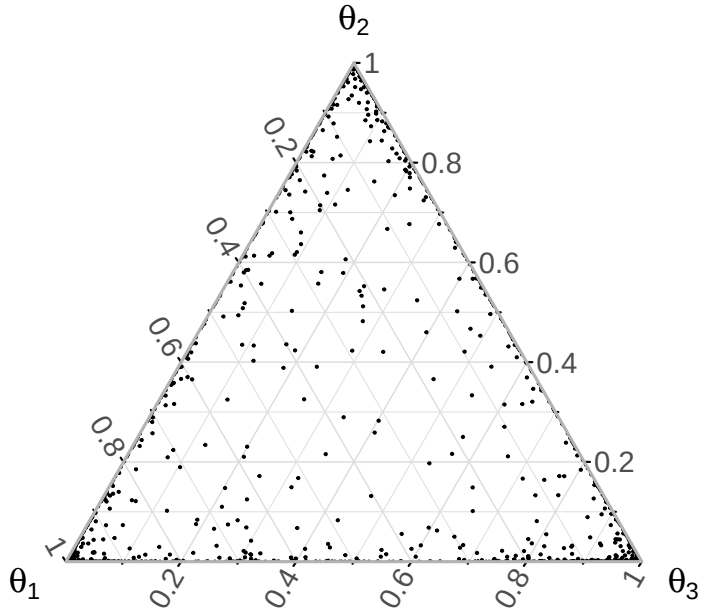


We see that the points are indeed uniformly spread across the triangle. Next let's check the Haldane limit, which corresponds to the improper distribution $p(\theta_1, \theta_2, \theta_3) \propto \theta_1^{-1}\theta_2^{-1}\theta_3^{-1}$. of course we can't literally take the parameters to zero, but we can make them small:

```
> alpha = c(0.1,0.1,0.1)
```

[13] Nicholas Hamilton (2016). ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams. R package version 2.2.0. `https://CRAN.R-project.org/package=ggtern`.

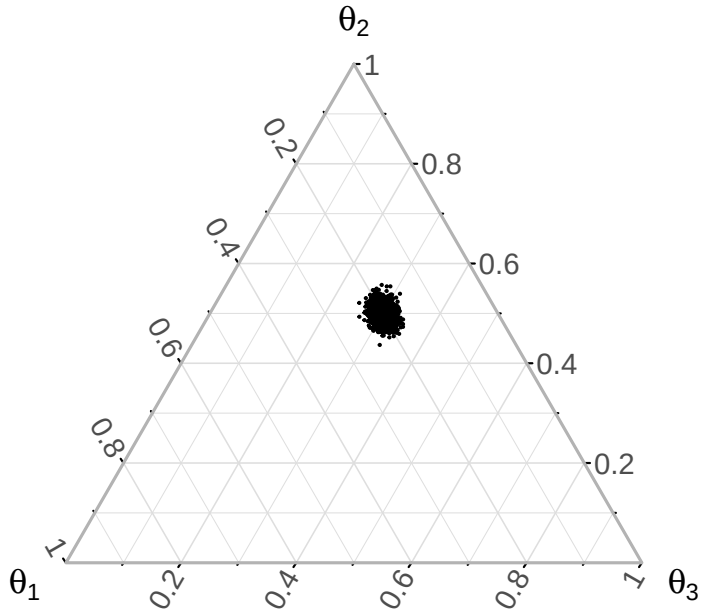
```
> thetaframe = myframe(rdirichlet(N,alpha))
> addmyopts(ggtern(thetaframe,aes(theta1,theta2,theta3)))
```



We see that the points tend to cluster on the edges and even more in the corners, which makes sense since the distribution function becomes large if as many θ_j as possible are close to zero.

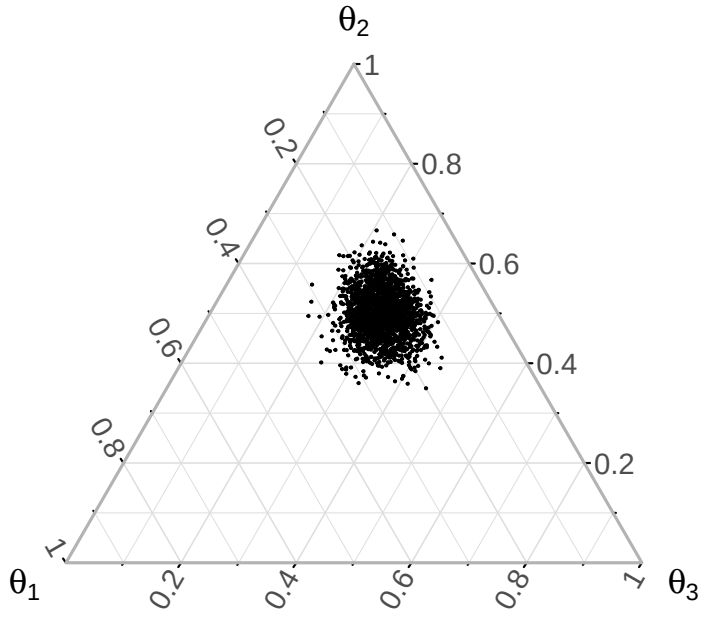
Now let's suppose we've seen some data with $\approx 20\%$ in the first category, $\approx 50\%$ in the second category, and $\approx 30\%$ in the third category. First we assume $n = 1000$, and see that we have a distribution peaked strongly at $\theta_1 = 0.20$, $\theta_2 = 0.50$, and $\theta_3 = 0.30$:

```
> alpha = c(200,500,300)
> thetaframe = myframe(rdirichlet(N,alpha))
> addmyopts(ggtern(thetaframe,aes(theta1,theta2,theta3)))
```



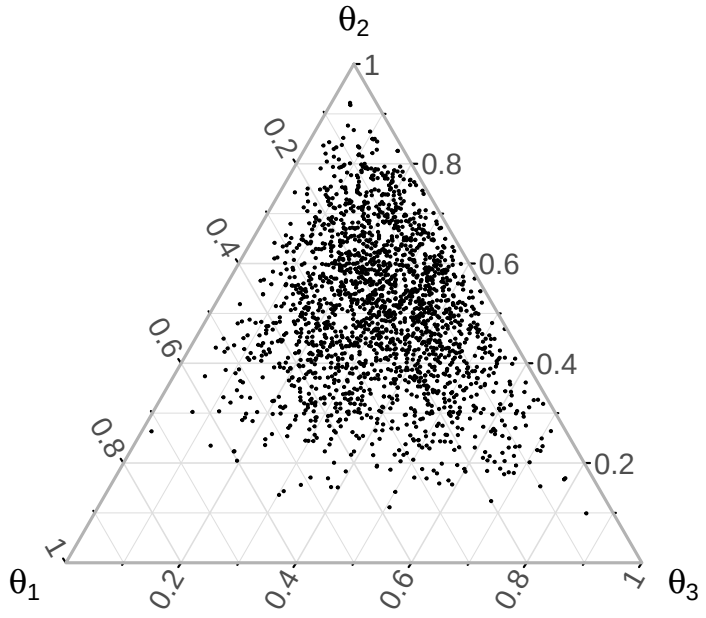
We can take $n = 100$, and we'll find the Dirichlet distribution is a little more spread out:

```
> alpha = c(20,50,30)
> thetaframe = myframe(rdirichlet(N,alpha))
> addmyopts(ggtern(thetaframe,aes(theta1,theta2,theta3)))
```



Finally, let $n = 10$, and we see that most of the allowed combinations of θ_j are still plausible. (Note that this is implicitly assuming we're starting with the Haldane prior. If we used the uniform prior, we'd get a Dirichlet distribution with $\boldsymbol{\alpha} = (3, 6, 4)$ rather than $(2, 5, 3)$. This is a significant difference when $n = 10$, but not so much when $n = 100$ or $n = 1000$.

```
> alpha = c(2,5,3)
> thetaframe = myframe(rdirichlet(N,alpha))
> addmyopts(ggtern(thetaframe,aes(theta1,theta2,theta3)))
>
```



3.4.3 Reparametrization

The parameters of a multinomial distribution when $J = 3$ provide a good illustration of how to handle a change of variables or reparametrization involving more than one parameter. Suppose we decide to consider not the probability parameters θ_1, θ_2, θ_3, but ϕ_1, ϕ_2, ϕ_3, where

- $\phi_3 = 1 - \theta_3$ is the probability that a given trial will *not* have result #3.
- $\phi_1 = \frac{\theta_1}{1-\theta_3}$ is the conditional probability that a given trial will have result #1, *if* it is known not to have result #3.
- $\phi_2 = \frac{\theta_2}{1-\theta_3}$ is the conditional probability that a given trial will have result #2, *if* it is known not to have result #3.

The question is, how do we go from a probability distribution like $p(\theta_1, \theta_2, \theta_3 | I_{\boldsymbol{\alpha}})$ to $p(\phi_1, \phi_2, \phi_3 | I_{\boldsymbol{\alpha}})$. It's not to hard to see that the conditions $0 < \theta_1, \theta_2, \theta_3 < 1$, $\theta_1 + \theta_2 + \theta_3 = 1$ correspond to

$0 < \phi_1, \phi_2, \phi_3 < 1$, $\phi_1 + \phi_2 = 1$. To transform the probability density, we note that we need to satisfy

$$p(\theta_1, \theta_2, \theta_3|I_{\boldsymbol{\alpha}})\, d\theta_1\, d\theta_2\, d\theta_3 = p(\phi_1, \phi_2, \phi_3|I_{\boldsymbol{\alpha}})\, d\phi_1\, d\phi_2\, d\phi_3 \quad (3.51)$$

where $d\theta_1\, d\theta_2\, d\theta_3$ can be thought of as the measure for a triple integral. We recall the result from multivariable calculus that

$$d\theta_1\, d\theta_2\, d\theta_3 = \left|\det\left\{\frac{\partial\theta_j}{\partial\phi_i}\right\}\right| d\phi_1\, d\phi_2\, d\phi_3 \quad (3.52)$$

where $\det\left\{\frac{\partial\theta_j}{\partial\phi_i}\right\}$ is known as the Jacobian determinant. To evaluate it, we need the inverse transformation

$$\theta_1 = \phi_1\phi_3 \quad (3.53a)$$
$$\theta_2 = \phi_2\phi_3 \quad (3.53b)$$
$$\theta_3 = 1 - \phi_3 \quad (3.53c)$$

which makes the Jacobian matrix

$$\left\{\frac{\partial\theta_j}{\partial\phi_i}\right\} = \begin{pmatrix} \frac{\partial\theta_1}{\partial\phi_1} & \frac{\partial\theta_1}{\partial\phi_2} & \frac{\partial\theta_1}{\partial\phi_3} \\ \frac{\partial\theta_2}{\partial\phi_1} & \frac{\partial\theta_2}{\partial\phi_2} & \frac{\partial\theta_2}{\partial\phi_3} \\ \frac{\partial\theta_3}{\partial\phi_1} & \frac{\partial\theta_3}{\partial\phi_2} & \frac{\partial\theta_3}{\partial\phi_3} \end{pmatrix} = \begin{pmatrix} \phi_3 & 0 & \phi_1 \\ 0 & \phi_3 & \phi_2 \\ 0 & 0 & -1 \end{pmatrix} \quad (3.54)$$

and its determinant $\det\left\{\frac{\partial\theta_j}{\partial\phi_i}\right\} = -\phi_3^2$. This means

$$d\theta_1\, d\theta_2\, d\theta_3 = \phi_3^2 d\phi_1\, d\phi_2\, d\phi_3 \quad (3.55)$$

and

$$\begin{aligned} p(\phi_1, \phi_2, \phi_3|I_{\boldsymbol{\alpha}}) &= \phi_3^2 p(\theta_1, \theta_2, \theta_3|I_{\boldsymbol{\alpha}}) \\ &= \phi_3^2[\phi_1\phi_3]^{\alpha_1-1}[\phi_2\phi_3]^{\alpha_2-1}(1-\phi_3)^{\alpha_3-1} \\ &= \phi_1^{\alpha_1-1}\phi_2^{\alpha_2-1}\phi_3^{\alpha_1+\alpha_2}(1-\phi_3)^{\alpha_3-1}; \\ &\qquad 0 < \phi_1, \phi_2, \phi_3 < 1,\ \phi_1 + \phi_2 = 1 \end{aligned} \quad (3.56)$$

So we see that the Dirichlet distribution on θ_1, θ_2, θ_3 corresponds to independent distributions for (ϕ_1, ϕ_2) and ϕ_3, the former being Dirichlet with parameters α_1 and α_2 and the latter being beta with parameters $\alpha_1 + \alpha_2 + 1$ and α_3.

Tuesday 21 February 2017

3.5 Multivariate Gaussian

A Linear Algebra: Reminders and Notation

If $\mathbf{A}$ is an $m \times n$ matrix:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \quad (A.1)$$

and $\mathbf{B}$ is an $n \times p$ matrix,

$$\mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix} \quad (A.2)$$

then their product $\mathbf{C} = \mathbf{AB}$ is an $m \times p$ matrix as shown in Figure 2 so that $C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$.

If $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$ is an $n \times m$ matrix with elements $B_{ij} = A_{ji}$:

$$\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nm} \end{pmatrix} = \mathbf{B} = \mathbf{A}^{\mathrm{T}} = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{mn} \end{pmatrix} \quad (A.4)$$

If $\mathbf{v}$ is an n-element column vector (which is an $n \times 1$ matrix) and $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{w} = \mathbf{Av}$ is an m-element column

$$\mathbf{C} = \mathbf{AB} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mp} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix}$$
$$= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} + \cdots + A_{1n}B_{n1} & A_{11}B_{12} + A_{12}B_{22} + \cdots + A_{1n}B_{n2} & \cdots & A_{11}B_{1p} + A_{12}B_{2p} + \cdots + A_{1n}B_{np} \\ A_{21}B_{11} + A_{22}B_{21} + \cdots + A_{2n}B_{n1} & A_{21}B_{12} + A_{22}B_{22} + \cdots + A_{2n}B_{n2} & \cdots & A_{21}B_{1p} + A_{22}B_{2p} + \cdots + A_{2n}B_{np} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}B_{11} + A_{m2}B_{21} + \cdots + A_{mn}B_{n1} & A_{m1}B_{12} + A_{m2}B_{22} + \cdots + A_{mn}B_{n2} & \cdots & A_{m1}B_{1p} + A_{m2}B_{2p} + \cdots + A_{mn}B_{np} \end{pmatrix} \tag{A.3}$$

Figure 2: Expansion of the product $\mathbf{C} = \mathbf{AB}$ to show $C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$.

vector (i.e., an $m \times 1$ matrix):

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \mathbf{w} = \mathbf{Av} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$
$$= \begin{pmatrix} A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n \\ A_{21}v_1 + A_{22}v_2 + \cdots + A_{2n}v_n \\ \vdots \\ A_{m1}v_1 + A_{m2}v_2 + \cdots + A_{mn}v_n \end{pmatrix} \tag{A.5}$$

so that $w_i = \sum_{j=1}^{n} A_{ij}v_j$.

If $\mathbf{u}$ is an n-element column vector, then $\mathbf{u}^{\mathrm{T}}$ is an n-element row vector (a $1 \times n$ matrix):

$$\mathbf{u}^{\mathrm{T}} = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} \tag{A.6}$$

If $\mathbf{u}$ and $\mathbf{v}$ are n-element column vectors, $\mathbf{u}^{\mathrm{T}}\mathbf{v}$ is a number, known as the *inner product*:

$$\mathbf{u}^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$
$$= u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{i=1}^{n} u_iv_i \tag{A.7}$$

If $\mathbf{v}$ is an m-element column vector, and $\mathbf{w}$ is an n-element

column vector, $\mathbf{A} = \mathbf{v}\mathbf{w}^{\mathrm{T}}$ is an $m \times n$ matrix

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} = \mathbf{A} = \mathbf{v}\mathbf{w}^{\mathrm{T}}$$
$$= \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \cdots & w_m \end{pmatrix} = \begin{pmatrix} v_1 w_1 & v_1 w_2 & \cdots & v_1 w_n \\ v_2 w_1 & v_2 w_2 & \cdots & v_2 w_n \\ \vdots & \vdots & \ddots & \vdots \\ v_m w_1 & v_m w_2 & \cdots & v_m w_n \end{pmatrix} \tag{A.8}$$

so that $A_{ij} = v_i w_j$.

If $\mathbf{M}$ and $\mathbf{N}$ are $n \times n$ matrices, the determinant $\det(\mathbf{MN}) = \det(\mathbf{M}) \det(\mathbf{N})$.

If $\mathbf{M}$ is an $n \times n$ matrix (known as a square matrix), the inverse matrix $\mathbf{M}^{-1}$ is defined by $\mathbf{M}^{-1}\mathbf{M} = \mathbf{1}_{n\times n} = \mathbf{M}\mathbf{M}^{-1}$ where $\mathbf{1}_{n\times n}$ is the identity matrix

$$\mathbf{1}_{n\times n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{A.9}$$

If $\mathbf{M}^{-1}$ exists, we say $\mathbf{M}$ is invertable.

If $\mathbf{M}$ is a real, symmetric $n \times n$ matrix, so that $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$, i.e., $M_{ji} = M_{ij}$, there is a set of n orthonormal *eigenvectors* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ with real eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, so that $\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Orthonormal means

$$\mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{A.10}$$

where we have introduced the Kronecker delta symbol δ_{ij}. The eigenvalue decomposition means

$$\mathbf{M} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{A.11}$$

The determinant is $\det(\mathbf{M}) = \prod_{i=1}^{n} \lambda_i$. If none of the eigenvalues $\{\lambda_i\}$ are zero, $\mathbf{M}$ is invertable, and the inverse matrix is

$$\mathbf{M}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{A.12}$$

If all of the eigenvalues $\{\lambda_i\}$ are positive, we say $\mathbf{M}$ is positive definite. If none of the eigenvalues $\{\lambda_i\}$ are negative, we say $\mathbf{M}$ is positive semi-definite. Note that these conditions are equivalent to the more common definition: $\mathbf{M}$ is positive definite if $\mathbf{v}^{\mathrm{T}}\mathbf{M}\mathbf{v} > 0$ for any non-zero n-element column vector $\mathbf{v}$ and positive semi-definite if $\mathbf{v}^{\mathrm{T}}\mathbf{M}\mathbf{v} \geq 0$ for any n-element column vector $\mathbf{v}$.